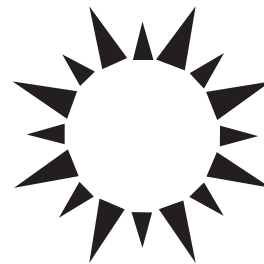


Philanthropy and Energy

ERIC HEITZ

The Energy Foundation
San Francisco, California, United States



-
1. Introduction to Energy and Philanthropy
 2. Qualities of Effective Philanthropy
 3. The Merits of Energy as a Philanthropic Endeavor
 4. Recent History: Energy and Philanthropy
 5. Conclusions: The Future Role for Foundations in Energy

Glossary

- energy/environmental nexus** The fact that the world's energy problems cause the majority of the world's environmental problems.
- foundation** A philanthropic organization; assets remain tax free as long as 5% of those assets are given for charitable and educational purposes each year.
- leverage** The ratio of philanthropic investment and the quantifiable change that results.
- philanthropy** (1) A desire to improve the material, social, and spiritual welfare of humanity; (2) a term applied to institutions that disperse wealth in pursuit of these goals.
- sustainable energy** The energy that is produced and used in ways that support long-term human development in all its social, economic, and environmental dimensions.
- Tax Reform Act of 1969** A reform of U.S. tax code, codifying the rules under which U.S. foundations operate.

Smart philanthropy can play a productive role in U.S. and global energy policy. It can help spur markets for the next generation of clean energy technologies that address the energy/environment nexus, especially global warming pollution. This article offers some principles for focused philanthropy in energy, discusses the merits of energy as a philanthropic endeavor, and outlines the recent history of energy and philanthropy in the United States.

1. INTRODUCTION TO ENERGY AND PHILANTHROPY

The best philanthropy involves the search for cause, an attempt to cure evils at their source.

—John D. Rockefeller, Sr.

Philanthropy's impact on society is both underappreciated and misunderstood. At their worst, the individual foundations that comprise the world's philanthropic sector are opaque institutions with vague goals; their funds disappear with no apparent trace. At their best, they are strategically driven, staffed by experts in the field, and can claim major contributions to the fields of human health and medicine, agricultural science, social policy, education, art, and, more recently, energy and the environment.

In essence, foundations are nonprofit institutions that disperse accumulated wealth. John D. Rockefeller and Andrew Carnegie, who built institutions to give away their considerable wealth at the end of the 19th century, are credited with birthing the modern grant-making foundation. The Tax Reform Act of 1969 codified their favorable tax treatment in the United States. In short, assets converted to foundation status are not taxed as long as 5% of those assets are given out each year for "charitable and educational purposes." Coupled with the growth in wealth in the last half of the 20th century, these rules spurred rapid growth of the foundation sector. Since 1975, the number of foundations has more than doubled and total giving (in constant dollars) has more than tripled. According to a report published by the American Association of Fundraising Counsel (AAFRC) Trust for Philanthropy, *Giving USA 2002*, over 56,000 registered foundations in the United States collectively gave out \$27.6 billion in funding. Figure 1 shows their investments. Of the total, only 7% was for environmental purposes, and

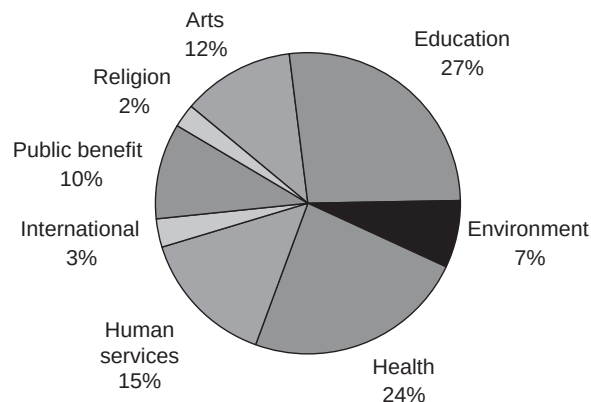


FIGURE 1 Philanthropic investment by field.

a fraction of that focused on energy. This investment pales in comparison to the nation's \$700 billion annual energy bill, raising the question, "What is the role for philanthropy in energy?" Some answers are proposed here, first by offering some characteristics of good philanthropy, followed by looking at the merits of energy as a philanthropic investment strategy, including some recent results, then concluding with some thoughts on the future role for philanthropy in energy.

2. QUALITIES OF EFFECTIVE PHILANTHROPY

Foundations do not have the resources to fund directly solutions to the world's energy challenges, whether that means transitioning to a clean energy future or simply providing basic energy services in developing countries. Focusing on the United States, where the challenge is to transition to a clean and low-carbon energy system, total foundation investments in energy pale in comparison to the nation's energy research and development budget of \$1.3 billion, which in turn is dwarfed by our annual energy bill. Given the scale of the investment flows compared to the size of the philanthropic resources, the theory of change adopted by all serious philanthropy work in energy is simple: public policy will change investment patterns, public or private, which will drive change (the U.S. experience is the focus here, but the same theory holds for international energy grant making). Since the 1970's, smart philanthropy has proved this approach, affecting policies that in turn change investment streams, often with a billion-to-one dollar or more leverage. Of course, it is impossible in a scientific way to attribute these

changes solely to foundation investments; political change is too complex. But careful students of the best philanthropy find remarkable returns, as is discussed in Section 4. So the question looms, "what makes good philanthropy?" Borrowing liberally from the writings of Hal Harvey, Director of the Hewlett Foundation's Environment Program, and Dr. Kenneth Prewitt, former Vice President of the Rockefeller Foundation, the following list of answers is offered:

1. Target root causes: To begin, it is essential, as John D. Rockefeller, Sr. advised, to target root causes. Alleviating symptoms will never change the underlying dynamic that created them; it may bring solace, but it will not speed change. This is why energy makes such an interesting philanthropic topic: it is the root cause of many environmental, economic, health, and international security problems facing the world.

2. Be focused: The biggest mistake many philanthropies make is to dabble in many areas instead of going deep in one area. This advice holds with respect to the choice of topics (e.g., marine conservation, forest preservation, energy) and choice of philanthropic means (e.g., research and analysis, policy development, policy advocacy, grassroots advocacy, or media work). It is very difficult to understand the political dynamics of any given sector. Deep insight is required to understand capital flows, to learn who the "brain trust" for a specific field is, and to learn the inclinations of the key decision makers. To do this in many fields at once is difficult. With respect to means, learning to fund the best research demands different skills, then choosing an effective grassroots advocate or trying to influence the media. The best philanthropies—those with a track record of measurable change—focus narrowly on a few topics and specialize in a few means.

3. Hire professional staff: To see the best prospect for change requires expertise equivalent to the best businesses or advocates in the field. The myth that giving away money is easy and thus any generalist can take up the task is one of the most harmful approaches to quality philanthropy.

4. Be strategy driven: Thinking through, writing down, and peer reviewing an explicit strategy of change is essential to good philanthropy. There are many paths to change. Philanthropies should know precisely what path they seek and what series of steps has to happen to affect change. Foundations should expose the premise of their strategy for debate and adjust their strategies, refining them to adapt to changing conditions as needed.

5. Scale to the problem: Solving large, complex problems, such as the nexus of energy/environment problems, begs for large philanthropic investments. Though the scale of philanthropic investments pales in comparison to the \$700 billion per year energy business, by focusing resources on key venues such as public utilities commission hearings or, in the transportation sector, air pollution authorities, foundation dollars can have a large impact.

6. Think long term: No serious energy or environmental problem can be solved overnight. Changing investment patterns, technologies, or consumer choices is a slow process, sometimes frustratingly so. Foundations, unlike most of the for-profit world, do not have next quarter's bottom line to define today's agenda. The best philanthropy takes the long view, investing for decades, not years.

3. THE MERITS OF ENERGY AS A PHILANTHROPIC ENDEAVOR

3.1 Energy as an Environmental and International Security Nexus

The goal of most philanthropy is to make the world better, whether improving conditions of the poor, stamping out disease, protecting endangered habitats, or increasing international security. Energy is interesting for philanthropy because it is simultaneously a key driver of economic growth and is at the center of Earth's most pressing environmental and security problems.

3.1.1 Environment

Fossil energy—coal and oil—has fueled tremendous economic growth over the past century, but at substantial, and growing, environmental cost. The U.S. Environmental Protection Agency (EPA) and Department of Energy (DOE) statistics show energy use is the leading cause of greenhouse gas emissions, smog, acid rain, air toxics, oil spills, and nuclear waste (Fig. 2). None of these problems can be solved without new energy trajectories. If the United States and the rest of world ramp up fossil fuel consumption to fuel further economic growth, there will be potentially catastrophic global environmental consequences. China, for example, is expected to surpass Western Europe's carbon emissions before 2010 and will surpass the United States to become the world's largest carbon emitter by about 2030. The Energy Information Administration (EIA) *International Energy Outlook 2003* report shows that Brazil and

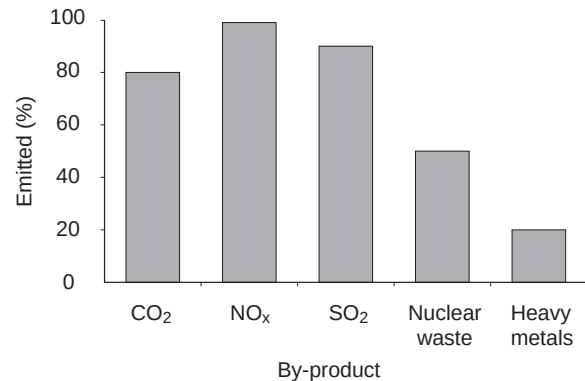


FIGURE 2 Percentage of emissions into the environment from energy use by-products.

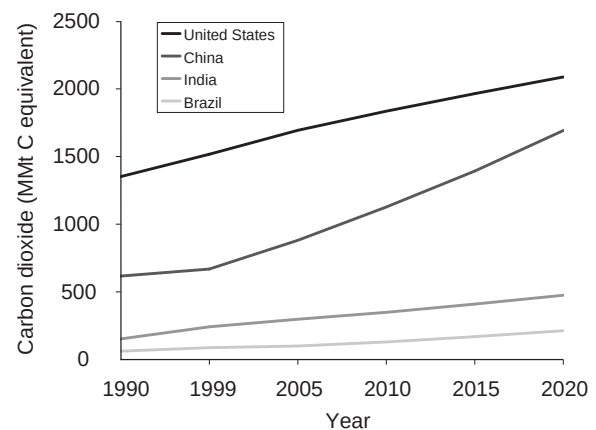


FIGURE 3 Projected carbon dioxide emissions. MMt C, million metric tons of carbon.

India are on similar trajectories for rapid fossil energy growth (Fig. 3).

The energy crises of the 1970s inspired great concern about running out of fossil fuels. Geology and technology have since revealed that there are vast reserves of coal, and large reserves of natural gas and oil. A major crisis in energy, instead, turns out to be that the capacity of the environment—our atmosphere, forests, soils, rivers, and oceans—to absorb the wastes of our energy-intensive society is threatened and will suffer long before energy supplies will be exhausted. *Laissez-faire* economics, wherein prices rise due to supply scarcity and so drive investment in new energy options, will not solve this dilemma. Sensible policy is needed to move developed and developing nations to a sustainable energy path. Foundations can help create and advance new policies.

3.1.2 International Security

The 2003 war in Iraq, a country situated on the world's second largest remaining oil reserves, brings

home the international security implications of the U.S. economy's dependence on imported oil. The world's largest remaining oil supplies and, even more important in the short term, the swing production that most influences world market price, are in the politically volatile Middle East. We import more oil today than at any time in U.S. history. Worse, the latest government forecast by the Energy Information Administration, *Annual Energy Outlook 2003*, projects a 60% increase in domestic oil use over the next 25 years, with imports rising to 67.8%. And some developing countries are on even faster oil consumption growth trajectories. China's automobile market, for example, is growing at almost 35% per year. In the International Energy Agency report, *China's Worldwide Quest for Energy Security*, projections show oil imports increasing to 8 million barrels per day (mbd) by 2020 (from 0 mbd in 1993), further increasing competitive pressures on diminishing world oil supplies.

3.2 Promising Technology Solutions

The compelling question is whether it is possible to build a benign energy system given a large and still-growing world population and universal aspirations for higher income. The answer to that question must be found in technology transformation. Our society needs to invent a low-carbon, low-pollution affluence. The energy field is replete with examples of how this might be done and, over the past several decades, examples of how philanthropy can help accelerate change. It is clear that technologies are available to reduce drastically pollution and carbon from buildings, industry, and transportation. Recent philanthropic efforts focus on catalyzing this technology transformation in the building sector, transportation field, and power sectors. There are promising trends in each sector.

3.2.1 Buildings

The case for technology transformation is perhaps easiest to make in the buildings sector. It is neither difficult nor costly to design a building that uses only nominal energy for heating and cooling and that requires only a quarter of the electricity of a conventional building. In the mid-1980s, California revamped its building codes toward this end. Title 24, as the resulting codes were named, reduces energy use in newly constructed buildings by some 75% compared to precode buildings. Appliance standards are another example of the transformative power of new technologies. Refrigerators sold today

are larger and have more features, yet use one-third the energy compared to 20 years ago. State appliance standards, eventually adopted by the federal government, drove these savings, with billion dollar savings for consumers. The American Council for an Energy Efficient Economy estimates that existing standards—for refrigerators, air conditioners, lights, washings machines, etc.—will save the equivalent of the energy generated by 400 large power plants, at net consumer savings of \$160 billion by 2020. Standards being contemplated for adoption by the federal government today could save consumers another \$75 billion dollars by 2020. Foundations have been the largest nongovernmental supporters of advancing such standards.

3.2.2 Transforming the Automobile

Transportation has the highest potential for technological advancement. The automobile is on the verge of a technological revolution. Today's automobiles are between 15 and 20% efficient; the lion's share of each gallon of gasoline goes directly out the exhaust pipe. Existing new technologies can double the efficiency of the engine, and double again the efficiency of the driveline, resulting in cars that exceed 100 miles per gallon with only modest air pollution (Fig. 4). High-efficiency tires, drivelines, and aerodynamic shells can reduce power requirements by a factor of two to four. Hybrid drives, which couple the efficiency of electric drives with the range and power output of liquid or gaseous fuels, can double engine/driveline efficiency. And fuel cells, which convert gas or liquid hydrogen directly into electricity at high efficiency with no moving parts and virtually no pollution, offer a very clean transportation future.

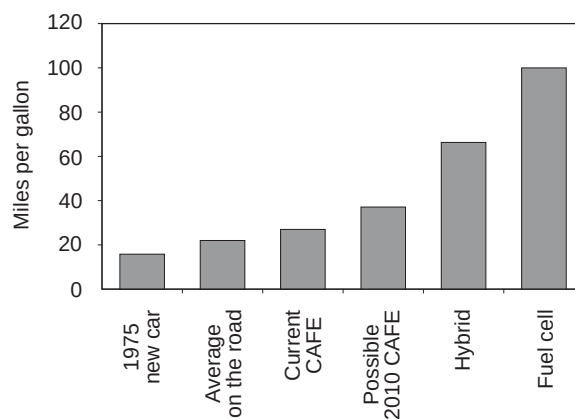


FIGURE 4 Automobile fuel efficiency. CAFE, Corporate average fuel economy (standards imposed under the Energy Policy and Conservation Act of 1975).

Toyota's Prius offers a hybrid drivetrain, including regenerative braking. The Prius gets almost 50 miles per gallon, but otherwise has features and performance similar to those of a Toyota Corolla. The Prius is exceeding all sales predictions, and Toyota recently doubled production volume. Honda has come out with a hybrid version of its popular Civic model, which gets 45 miles per gallon. Ford expects to introduce a sport utility vehicle (SUV) hybrid that is projected to obtain 40 miles per gallon, more than double today's average SUV mileage. Toyota is expected to offer an SUV on its Lexus 330 platform as well. GM has also announced hybrid versions. Foundations have a role in making the analytic case for these technologies and in spurring the policies that spur their commercialization.

3.2.3 Cleaner Power

Renewable energy technologies have enjoyed dramatic price reductions over the past two decades. Wind power has dropped from 25¢/kilowatt-hour to less than 3¢/kWh today. The central region of the United States is known as the "Saudi Arabia of wind energy." The 11 contiguous states from the Dakotas to Texas have the wind energy potential to supply more than the total current electricity consumption of the entire United States (the states are Montana, Wyoming, North Dakota, South Dakota, Minnesota, Iowa, Nebraska, Kansas, Colorado, Oklahoma, and Texas). Driven largely by state policies spurring wind investments—which foundation-supported advocates championed in many states—wind is the fastest growing source of generation in the United States. California, the world's fifth largest economy, has adopted a goal of 20% renewable power generation by 2015. This policy will spur nearly \$10 billion in new wind investments. New York has announced a goal of 25%

renewable energy by 2010, and 12 other states have similar renewable energy goals, which are driving new markets for renewable energy developers (Fig. 5).

4. RECENT HISTORY: ENERGY AND PHILANTHROPY

4.1 The Ford Foundation Energy Policy Project

Major foundation investments in energy began in the 1970s, driven partly by the emerging environmental movement and more directly by the Organization of Petroleum Exporting Countries (OPEC) oil crisis. Under McGeorge Bundy, former National Security Advisor to Presidents Kennedy and Johnson, The Ford Foundation launched The Energy Policy Project in 1971, ultimately spending \$4 million on a series of energy studies. The most influential study, published in 1974, was *A Time to Choose*, by S. David Freeman. It was pathbreaking in two respects. First, it argued that economic growth and energy demands need not grow in lockstep—conservation and efficiency could decouple them. Second, it laid out the policies—from auto fuel efficiency regulations, to appliance efficiency standards, to federal energy efficiency research and development—that continue to be the mainstays of the U.S. energy efficiency policy today.

The Ford Foundation followed *Time to Choose* with two other major studies. In early 1979, they released *Nuclear Power Issues and Choices*, which concluded that, despite its many problems, nuclear energy should be a major power source in the future. Most important today, it argued that plutonium processing in breeder reactors was both uneconomic and created too great a danger of plutonium prolifera-

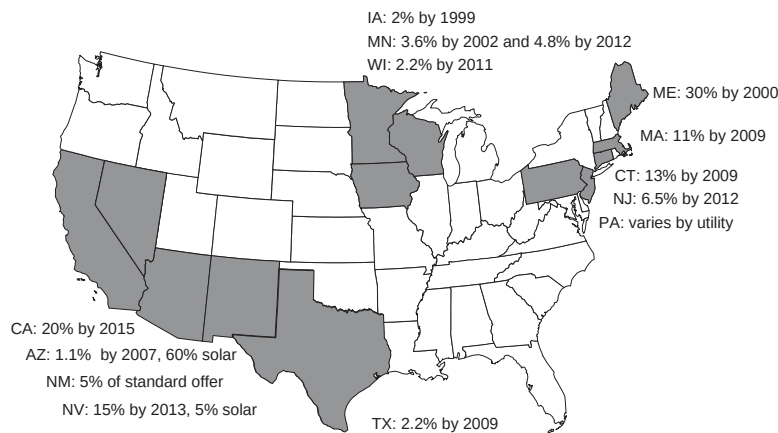


FIGURE 5 State renewable energy policies.

tion. One of the lead authors of the study, Dr. Spurgeon Keeny, Jr., argues that the study was a major factor in President Jimmy Carter's decision to defer commercial reprocessing of plutonium indefinitely. A third Ford Foundation study, released in 1979, *Energy, the Next Twenty Years*, was the least influential of the three studies; it concluded that the energy problem was essentially economic in nature. The U.S. would not run out of energy (given coal and oil shale reserves) any time soon, but prices would inevitably rise. Its main conclusion, that society would benefit from having the price of energy reflect its true value, continues to be the holy grail of energy economists today.

4.2 The Next Wave: From Studies to Advocacy

The premise of The Ford Foundation's investments in the 1970s was that serious new policy studies—the equivalent of investing \$14 million today—would capture high-level decision-maker attention and move the country along a more “intelligent” energy path. As the decade progressed, and as the environmental movement matured, a handful of foundations began to invest based on a different premise—namely, that direct advocacy was needed to change policies, which, in turn, would shape energy investments for the country. The crucible in which these early advocacy efforts were born was the nuclear power debate. Small, environmentally focused foundations such as the Mertz Gilmore Foundation, Levinson Foundation, Nathan Cummings Foundation, New-Land Foundation, List Foundation, J. M. Kaplan Fund, Compton Foundation, and the Rockefeller Family & Associates supported technical advocacy in state regulatory proceedings to (1) attack nuclear industry cost and safety claims and (2) make the case for cheaper and cleaner alternatives, particularly natural gas power plants and energy efficiency. Historians attribute the decline of the nuclear industry to many factors, including cost overruns, the public movement against nuclear power, and the fear of accidents such as the 1979 Three Mile Island incident. The early foundation-funded advocacy efforts deserve some of the credit as well. They also spawned a new approach to foundation-supported advocacy that continues to this day. The leaders in the energy policy advocacy work included the Mertz Gilmore Foundation, an early supporter of global warming policy work and one of the first foundations to offer grants in support of renewable energy. Other foundations taking on energy issues at this time were the Jessie Smith Noyes Foundation, the Evergreen Fund, the Changing

Horizons Charitable Trust, the Conservation and Research Foundation, and the Aria Foundation.

Following on the heels of Earth Day 1990 and the Gulf War in 1990/1991, which put global warming and oil dependency in the public eye, the presidents of the John D. and Catherine T. MacArthur Foundation, the Rockefeller Foundation, and the Pew Charitable Trusts launched a special-purpose foundation focused entirely on the energy/environmental nexus. Opening its doors in January 1991, the Energy Foundation received a \$100 million commitment for its first decade. Focused entirely on advancing energy efficiency and renewable energy, the Energy Foundation's first decade concentrated primarily on advancing state policies to spur commercialization of clean energy technologies. In the utility sector, for example, foundation-supported advocates in a dozen states advanced policies to make energy efficiency investments for utilities as profitable as building power plants. By 1994, U.S. utility investments in energy efficiency nearly tripled to \$2.7 billion per year, halving growth in electric demand. Even after a major restructuring of the utility industry, which slashed all capital investments by utilities, spending on energy efficiency by utilities today exceeds \$1 billion per year.

The Energy Foundation also invested heavily in state efforts to advance the cleanest and most efficient cars. Beginning in California in the early 1990s, and moving to New York, Massachusetts, Maine, and Vermont, scientists, energy experts, and grassroots groups worked in state air regulatory proceedings, arguing that only the most advanced vehicle technologies would clean the air adequately to meet public demands for healthy air. Because California, New York, Massachusetts, Maine, and Vermont represent 20% of the U.S. automobile and light-truck market, new tailpipe requirements captured the attention of the automobile industry. California's low-emission vehicle and zero-emission vehicle standards, and the subsequent adoption of the standards by New York and Massachusetts, are credited by many experts as the primary drivers spurring new investments in hybrid technology and fuel cell vehicles. Today, the Toyota Prius and Honda Civic hybrids are five-passenger production cars that are superclean and get nearly 50 mpg in real-world driving. Toyota's goal is to sell 300,000 hybrids annually worldwide by the end of this decade. GM, Ford, and Chrysler have all announced plans for hybrid vehicles as well. The newly revised California Air Resources Board rules, released April 2002, called for minimum annual sales of 110,000 hybrids by 2009.

These two examples—working toward reduction of electric demand and advocating for fuel efficiency—are indicative of Energy Foundation work. The foundation continues to invest in the power sector to advance efficiency and renewable energy, to promote superefficient building technologies, and to spur commercialization of advanced-technology vehicles. The Energy Foundation has evolved into a project of major foundations advancing clean energy technologies. In 1996, the Mertz Gilmore Foundation joined, and in 1998, the McKnight Foundation joined. The David and Lucile Packard Foundation joined in 1999, launching the China Sustainable Energy Program, which helps China promote energy efficiency and renewable energy technologies and employs nine Chinese nationals in a Beijing office. The William and Flora Hewlett Foundation joined in 2002, supporting advanced vehicle technology work in China and the United States as well as work to advance clean energy in the western United States. The Energy Foundation annual budget is now around \$20 million per year.

The Energy Foundation is the only special-purpose energy philanthropy in the United States, but it is by no means alone in supporting energy work. Under its Environment Program, the Pew Charitable Trusts has been a major investor making the case for action on global warming emissions. At the federal level, this includes the Pew Center on Global Climate Change (www.pewclimate.org), which built a coalition of major businesses that endorse action on climate and which has been a player in international deliberations and U.S. policy debates. The Pew Charitable Trusts, independent of the Energy Foundation, also built a state network of advocates focused on the health and environmental damages from the many Eisenhower-era coal power plants that still generate a significant share of U.S. electric power. Many of these power plants—in Texas, Ohio, Minnesota, and North Carolina—have been cleaned up or shut down as a result, and the federal debate to cap power sector emissions and, possibly, control carbon emissions from power plants is active. The William and Flora Hewlett Foundation is also a major new investor in energy work, beyond its investment in the Energy Foundation. Hewlett is focused on advanced vehicle technology in Mexico and Brazil, as well as on energy production on western U.S. lands. The Turner Foundation has also made energy investments, as have the Rockefeller Brothers' Fund and the W. Alton Jones Foundation. Major regional funders have also invested in energy, especially the Joyce Foundation in the midwestern United States, the John Merck Fund in New England, and the Bullitt Foundation in the Pacific northwest,

among others. This author estimates that annual philanthropic support for energy (broadly defined to include global warming work) in the United States is about \$40 million per year in 2003.

5. CONCLUSIONS: THE FUTURE ROLE FOR FOUNDATIONS IN ENERGY

Foundations have made good progress on the energy front, but huge gaps remain. Promising technologies in use and production today prove that the prospects for an affluent but low-carbon society are good. But the U.S. and the world economies continue to run primarily on oil and coal. As the largest consumer and producer of energy in the world, the United States has the responsibility to lead the transition to cleaner technologies. Philanthropy can play a role in this transition by spurring policies that commercialize the best technologies.

But U.S. leadership is not enough. Rapidly developing economies, such as those of China, India, and Brazil, cannot afford to follow our path of energy consumption—they need opportunities to leapfrog to advanced technologies that offer higher standards of living but with radically reduced energy use. Here, too, philanthropic opportunities in energy abound, with potential billion-to-one payoffs. In essence, we are confronted with the ultimate “diffusion of innovations” challenge. Technology learning curves—for clean technologies such as wind power, photovoltaics, or superefficient gas turbines—tell us that each cumulative doubling of production can reduce prices by 20–30%. In turn, lower prices speed the penetration of these new energy technologies into U.S. and world markets. The history of the past three decades suggests energy policy, in the absence of higher price signals driven by energy scarcity, will be the major factor in how fast we move down these clean technology cost curves and toward a transformed energy economy. They also suggest that smart philanthropy can play a productive role.

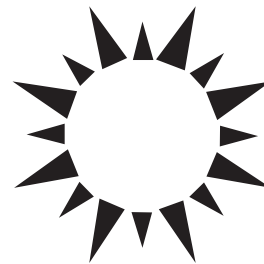
SEE ALSO THE FOLLOWING ARTICLES

Equity and Distribution in Energy Policy • Non-governmental Organizations (NGOs) and Energy • United Nations Energy Agreements • World Environment Summits: The Role of Energy

Further Reading

- Center on Philanthropy at Indiana University. (2002). "Giving USA 2002." American Association of Fundraising Counsel (AAFRC) Trust for Philanthropy, Indianapolis.
- Energy Foundation. (2004). Available on the Internet at <http://www.ef.org/>.
- Environmental Grantmaking Foundations. (2003). "Environmental Grantmaking Foundations," 2003 ed. Resources for Global Sustainability, Inc., Cary, North Carolina.
- Harvey, H. (1999). The importance of focus in effective philanthropy. *In* "New Directions for Philanthropic Fundraising, The Changing World of Foundation Fundraising: New Challenges and Opportunities" (S. A. Glass, Ed.). Jossey-Bass, San Francisco. Chap. 2.
- Nielsen, W. A. (1996). "Inside American Philanthropy: The Dramas of Donorship." University of Oklahoma Press, Norman.
- Prewitt, K. (1987). Foundations. *In* "The Nonprofit Sector: A Research Handbook," (W. W. Powell, Ed.). Yale University Press, New Haven Publication forthcoming.
- Prewitt, K. (1995). "Social Sciences and Private Philanthropy: The Quest for Social Relevance. Series on Foundations and Their Role in American Life. 15." Indiana University Center on Philanthropy, Indianapolis.

Photosynthesis and Autotrophic Energy Flows



JED P. SPARKS

Cornell University
Ithaca, New York, United States

-
1. Introduction
 2. The Photosynthetic Process
 3. Chemoautotrophy
 4. Photorespiration
 5. Photosynthetic Pathway Variation
 6. Global Carbon Cycling: Photosynthetic Considerations
 7. Global Climate Implications
 8. Conclusion

Glossary

adenosine triphosphate (ATP) Energy-bearing molecule formed during the light reactions through the phosphorylation of adenosine diphosphate. The energy bound in ATP is used as the primary energy source for most biological reactions.

C₄ photosynthesis In order to reduce the amount of carbon lost during photorespiration, some plants have developed C₄ photosynthesis. These plants initially carboxylate the acceptor molecule phosphoenolpyruvate to form four-carbon organic acids (hence the term C₄ photosynthesis). These molecules are transported internally to specialized bundle sheath cells. There, the plant enzymatically breaks down the acid and releases CO₂. The CO₂ is then fixed again using the normal enzymes of the dark reactions. As a result, the concentration of CO₂ at the point of the dark reactions is much higher than that of the surrounding atmosphere and effectively eliminates photorespiration.

chemoautotrophs Organisms that use CO₂ as a source of carbon, but rather than using light energy they obtain reducing power and energy from the oxidation of inorganic compounds.

dark reactions The process of using the reducing power and energy produced during the light reactions to fix inorganic material into high-energy organic molecules.

light reactions The process of coupling light energy to the creation of chemical reducing power and energy.

This process occurs in the thylakoid membrane in higher plants.

nicotinamide adenine dinucleotide phosphate (NADPH) Reducing molecule formed during the light reactions. Reducing energy of NADPH is used to reduce (i.e., store energy) inorganic material making organic high-energy molecules.

oxidation–reduction reaction Chemical reaction involving the transfer of electrons. Reduction is the transfer of an electron from a donor molecule to an acceptor molecule. During the reaction, the donor is said to become oxidized and the acceptor reduced.

photorespiration The RuBP carboxylase–oxygenase enzyme also exhibits oxygenase activity (i.e., it will catalyze the reaction of O₂ with RuBP). This leads to the loss of a CO₂ molecule previously fixed and decreases the efficiency of photosynthetic carbon assimilation.

photosynthesis The process of converting light energy to organic chemical energy.

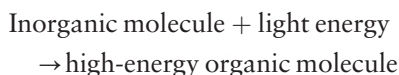
RuBP carboxylase oxidase Arguably one of the most important enzymes on Earth. RuBP carboxylase oxidase (or Rubisco) catalyzes the reaction of CO₂ and an acceptor molecule, ribulose biphosphate (RuBP), to form 3-phosphoglyceric acid (3-PGA). This is the initial carboxylation step of the dark reactions of photosynthesis.

Photosynthesis and the resultant autotrophic energy flow is a multistep pathway that is arguably the most important set of chemical reactions on earth. The reduction of compounds (i.e., the storing of energy in chemical bonds) is the foundation of all biological metabolism. Photosynthesis and, to a lesser extent, chemoautotrophy are the processes of this reduction. The capture of light energy or the oxidation of inorganic material produces energy and reducing power to form high-energy molecules necessary for life. This article gives a general introduction and overview of photo- and chemoautotrophic reactions and the organisms who perform these reactions, and

it briefly outlines the recent human impact on photosynthetic processes and global carbon cycling.

1. INTRODUCTION

Photosynthesis or photoautotrophism is the mechanism of converting light energy to organic chemical energy performed by certain bacteria, algae, and plants. The photosynthetic process consists of both physical and chemical reactions enabling organisms to exploit an energy resource external to the Earth's system: the sun. The organic molecules produced by these organisms provide the energy powering nearly all living organisms. In short, photosynthetic reactions directly or indirectly provide energy for most biological reactions and may be expressed as:



According to the laws of thermodynamics, complex systems will tend to move from a state of organization (i.e., high energy) to a random state (i.e., low energy) through time until equilibrium is reached. In the case of living organisms, equilibrium between the organism and the environment is the death and decomposition of the organism. Therefore, to stay alive an organism must consume energy to reverse the thermodynamic direction. Living organisms are not in balance in terms of thermodynamics and require energy to maintain structure, accumulate matter, conduct metabolism, grow, and, eventually, reproduce.

All nonphotosynthetic organisms, such as animals, fungi, and bacteria, are dependent on organic materials originally generated through photosynthetic (or chemosynthetic) reactions. Without this supply of high-energy molecules, organisms would not survive. Only those organisms able to utilize the supply of energy from the sun increase the total free energy of the earth and increase the energy available to organisms unable to use either chemical or solar energy sources, including humans.

The photosynthetic process is found in some bacteria, cyanobacteria (often referred to as the blue-green algae), algae, and higher plants. Given a supply of sunlight, photosynthetic organisms can survive, grow, and reproduce using only inorganic forms of matter. Molecules, known as pigments, that absorb light and convert physical energy to chemical energy have allowed the evolution and proliferation

of complex multicellular organisms, dependent on light as their ultimate energy source.

2. THE PHOTOSYNTHETIC PROCESS

In simple terms, photosynthesis is the process by which light energy is used to convert chemical substances to a more energy-filled state. In all photosynthetic organisms, it involves the capture of the energy of a photon of light by a pigment molecule, the formation of an excited electron, and the use of the excited electron to reduce an acceptor substance and to form reduced or energy-rich molecules (Fig. 1). Photosynthesis is often thought of as a two-stage process. The first stage, the light reactions, is the process of coupling light energy to the creation of chemical reducing power and energy. The second stage, the dark reactions, is the process of using the reducing power and energy produced during the light reactions to fix inorganic molecules into organic molecules. This two-stage process is outlined in detail in the following sections.

2.1 Oxidation–Reduction Reactions

The foundations of photosynthesis are a series of oxidation and reduction reactions. Very simply defined, reduction is the transfer of an electron (e^-) from a donor (M_d) molecule to an acceptor molecule

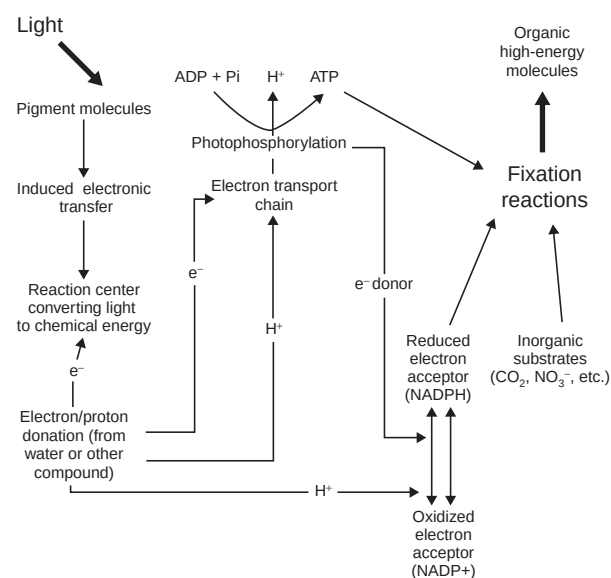
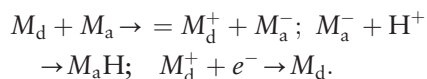


FIGURE 1 The basic reactions found in most photosynthetic organisms.

(M_a). During the reaction, the donor is oxidized and the acceptor is reduced. An electrically neutral compound becomes negatively charged and will extract a proton from water to restore electrical neutrality:



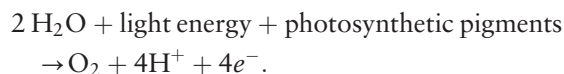
Reduction and oxidation reactions are fundamental to the photosynthetic process. The primary reaction of photosynthesis is the linking of light energy to a biochemical process. This is accomplished by the transfer of electrons from a light-excited chlorophyll pigment to an acceptor molecule. The acceptor molecule (in higher plants, a quinone) is reduced and the chlorophyll oxidized. Electrons are then donated from various sources (water in the case of higher plants) to the oxidized chlorophyll, reducing it and allowing the process to repeat.

The separation of the photosynthetic process into the light and dark reactions is primarily a historical artifact and has led to confusion regarding the use of terms. The generation of both the reducing [nicotinamide adenine dinucleotide phosphate (NADPH)] and the energy-containing molecules [adenosine triphosphate (ATP)] requires light energy, and these are logically referred to as the light reactions. The NADPH and ATP molecules are subsequently used to reduce other inorganic substances and convert them to organic molecules. With an adequate supply of NADPH and ATP, the chemical reactions of photosynthesis will proceed irrespective of light. The dark reactions of photosynthesis can and do occur in the light and darkness.

2.2 Capture of Light Energy and Electron Donation: The Light Reactions

Pigment molecules of various kinds allow photosynthetic organisms to both capture light and convert the captured light energy into chemical energy. Light-harvesting pigments (primarily the chlorophylls) capture light from the sun and form excited pigments. The combined excitation energy of multiple pigment molecules is focused in a single pigment molecule known as the reaction center (RC) and used to reduce acceptor molecules. The light-harvesting apparatus (the RC plus multiple accessory pigments) or antenna pigments absorb light and excitation energy is produced. This excitation energy moves through the antenna molecules to the RC where an electron is excited and lost. The electron that is lost from the pigment is then replaced by an electron

from another source in the environment. In photosynthetic bacteria, a wide variety of compounds may donate electrons to the oxidized pigment. In cyanobacteria, algae, bryophytes, ferns, and higher plants water is used as the source of electrons and diatomic oxygen (O_2) is released:

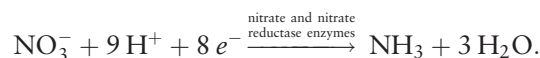


The most common pigment molecule in higher plants is chlorophyll, but there are many other pigments involved in photosynthesis across different organisms, including phycocyanin, phycococerythrin, and allophycocyanin in blue-green and red algae and carotenoids and xanthophylls in most higher plants and some algae and bacteria.

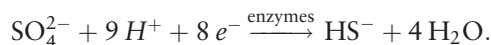
In both bacteria and higher photosynthetic organisms the excited electrons are used to form reducing molecules, either nicotinamide adenine dinucleotide (NADH) or, more commonly in photosynthetic organisms, NADPH. Simultaneously, through electron transport, protons accumulate on one side of a cellular membrane, forming a proton concentration gradient, the energy of which is used to produce ATP through the phosphorylation of adenosine diphosphate. ATP and NADPH provide reducing power and energy to nearly all biological energy and matter transformation reactions.

2.3 Reduction of Inorganic Material: The Dark Reactions

All organic biomass created by photosynthetic organisms is generated using reductant (NADPH) and energy (ATP) from the light reactions. This energy (i.e., ATP and NADPH) of photosynthesis may eventually be used to reduce a range of organic substances. The most common compound reduced is carbon dioxide (CO_2), leading to the synthesis of common sugars (CH_2O) and other organic molecules based on carbon. The fixation of carbon is the focus of this section. However, ATP and NADPH from the light reactions are also used to assimilate other inorganic compounds. Nitrate ions (NO_3^-) are reduced to ammonia, which is synthesized to amino acids and, eventually, proteins:

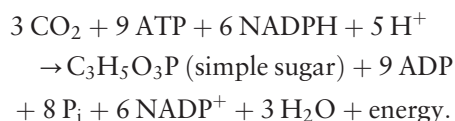


Sulfate ions must also be reduced before entering biological metabolism:



Photosynthesis, although almost always associated with the fixation of carbon, is a process with many possible products and it participates in many facets of metabolism.

The dark reactions of photosynthesis in higher plants occur within the chloroplast. The rate of CO₂ fixation is determined by the enzyme activity, the supply of CO₂, and the conditions within the chloroplast. The assimilation of carbon dioxide is a cyclical process. The assimilation (i.e., fixation or reduction) of CO₂ can be simply described as



This scheme is a simplification listing only the substrates and energy inputs. The process includes 11 enzymes catalyzing a 13-step carboxylation cycle. In the initial carboxylation reaction, the enzyme ribulose biphosphate (RuBP) carboxylase catalyzes the reaction of CO₂ and an acceptor molecule, RuBP, to form 3-phosphoglyceric acid (3-PGA). RuBP is then regenerated from 3-PGA in reactions consuming NADPH and ATP. For each CO₂ assimilated, a minimum of three ATP and two NADPH are needed from the light reactions. The roles of the dark reactions are therefore to regenerate RuBP for additional CO₂ assimilation and to produce simple sugars.

Interestingly, the enzyme RuBP carboxylase exhibits dual activity. RuBP carboxylase will catalyze CO₂ assimilation as described previously or the assimilation of O₂—oxygenation. Because of this dual functionality, the enzyme is often referred to as RuBP carboxylase oxygenase or Rubisco. The ability of Rubisco to catalyze oxygenation reactions has implications for photosynthetic energy flows and will be considered later.

2.4 Variations in Photosynthesis among Biological Organisms

The most primitive type of photosynthesis is thought to be that exhibited in the halophilic bacterium, *Halobacterium halobium*. These organisms have patches of the pigment bacteriorhodopsin on their outer membrane. This pigment captures light and transports H⁺ from inside the cell to the outside. This process forms a pH/electrical gradient that the organism then uses to synthesize ATP when the H⁺ diffuses back into the cell. This is the simplest form of photosynthetic reaction and is not associated with

more complex electron transport systems found in other photosynthetic organisms.

Most photosynthetic bacteria exhibit a slightly more complex photosynthetic system in which electron transport is used to form a pH gradient for ATP synthesis. An electron transport chain greatly increases the efficiency of ATP synthesis and allows for the generation of NADPH. To facilitate an electron transport chain, the outer cell membrane of most photosynthetic bacteria is folded, creating an enclosed space where a chemical gradient can be generated between the pocket and the interior of the cell. Within these pockets, bacterial chlorophyll and associated pigments bound in the membrane capture light and form a reduced acceptor. Simultaneously, the transported electrons reduce quinone compounds within the membrane that transport H⁺ from the interior of the cell to the pockets. A pH gradient is created and the diffusion of H⁺ back into the cell drives the synthesis of ATP. In contrast to *Halobacterium*, these more complex photosynthetic bacteria utilize electron transport and produce NADH in addition to ATP.

The prokaryotic cyanobacteria (blue-green algae) produce oxygen during photosynthesis similar to higher plants and have a membrane structure similar to that of the complex bacteria: extensive membranes arranged around the exterior of the cell in sheets. The red algae are also similar to the complex bacteria. The membranes have regularly arranged patches of pigment called phycobilisomes. The phycobilisomes contain a grouping of pigments linked to a single reaction center. There are two types of complexes in the membranes, one associated with electron transport (photosystem I) and the other associated with water oxidation and electron further transport (photosystem II). As in the bacteria, transport of electrons leads to the synthesis of reductant (NADPH) and the formation of a pH gradient, which drives the synthesis of ATP.

In higher algae, bryophytes, ferns, and higher plants, the photosynthetic membranes (thylakoids) are enclosed within a double-membraned organelle known as a chloroplast. Multiple chloroplasts can be found within a single plant cell. The inside of the thylakoid membrane is called the lumen and is chemically equivalent to the outside of the cell. There are two photosystems, as in the cyanobacteria, but the higher photosynthetic organisms use some form of chlorophyll as the primary pigment and a variety of secondary pigments. Also, as in the cyanobacteria, electron transport leads to the reduction of an acceptor, the formation of a pH gradient, and the synthesis of ATP.

3. CHEMOAUTOTROPHY

Not all organisms require light to reduce inorganic molecules. These organisms are referred to as chemoautotrophs and are found entirely within the bacteria. Chemoautotrophy has received considerable attention since the discovery of several deep-sea vent communities in the Galapagos rift in 1977. The areas surrounding these sea vents contain entire communities based on energy from chemoautotrophic bacteria; primarily sulfur oxidizers. These bacteria use CO_2 as a source of carbon and obtain reducing power and energy from the oxidation of elemental sulfur, hydrogen sulfide, or thiosulfite. The submarine volcanic vents with which these organisms are associated discharge large quantities of sulfide-rich warm water. Although unique in the sense of an entire community based on chemoautotrophically derived energy, deep-sea vent communities are not the only environments to support chemoautotrophic bacteria. Other sulfur-oxidizing bacteria are known in deep freshwater lakes, in surface springs, and in caves. In addition, other chemosynthetic bacteria are known to oxidize ammonium (NH_4^+), nitrite (NO_2^-), iron (Fe^{2+}), hydrogen (H_2), or carbon monoxide (CO). Of these, the nitrifying bacteria, which oxidize ammonium to nitrite and nitrite to nitrate, are among the most ecologically important organisms on Earth due to their role in the cycling of global nitrogen between organic and inorganic pools. Without them, much of the nitrogen necessary for plant growth would be unavailable.

4. PHOTORESPIRATION

The oxygenase activity of the RuBP carboxylase-oxygenase enzyme is significant because it leads to the loss of a CO_2 molecule previously fixed and decreases the efficiency of photosynthetic carbon assimilation. When RuBP catalyzes an oxygenase reaction, it produces a phosphoglycolate (PG) and 3-PGA. The PG then moves out of the chloroplast and is metabolized by the glycolate pathway and leads to the eventual release of CO_2 . This is a form of respiration to the plant and offsets the amount of carbon fixed during assimilation. This process occurs during the light and is directly associated with photosynthesis; it is therefore referred to as photorespiration. The amount of photorespiration (i.e., the affinity of Rubisco for oxygen as a substrate) is dependent on temperature and the ratio of O_2 to

CO_2 in the chloroplast. Early in Earth's history when the photosynthetic reactions were thought to have evolved, the ratio of O_2 to CO_2 concentration was relatively low compared to today. Under these conditions, photorespiration is negligible. However, in the current atmosphere, in which the partial pressure of O_2 is 1000 times higher than that of CO_2 , 20–30% of the fixed carbon may be lost to photorespiration. Some plants have developed mechanisms that elevate the concentration of CO_2 inside the leaves, thereby decreasing or eliminating photorespiration. These specialized mechanisms are discussed in the following section.

5. PHOTOSYNTHETIC PATHWAY VARIATION

To minimize the effects of photorespiration, some plants have developed anatomical and biochemical characteristics allowing them to concentrate CO_2 at the site of Rubisco activity. The traditional photosynthetic dark reactions are often referred to as C_3 photosynthesis because the initial product of carboxylation is the three-carbon compound, 3-PGA. Some plants in hot and dry environments, in which photorespiratory loss of carbon is very high, have developed an alternative pathway, the C_4 pathway, to minimize carbon loss. The internal leaf anatomy and biochemistry of a C_4 plant is significantly different than that of a C_3 plant (Fig. 2). C_3 plants have vascular bundles surrounded by a layer or sheath of large colorless cells. Other cells, known as palisade cells, contain chlorophyll and are arranged against the

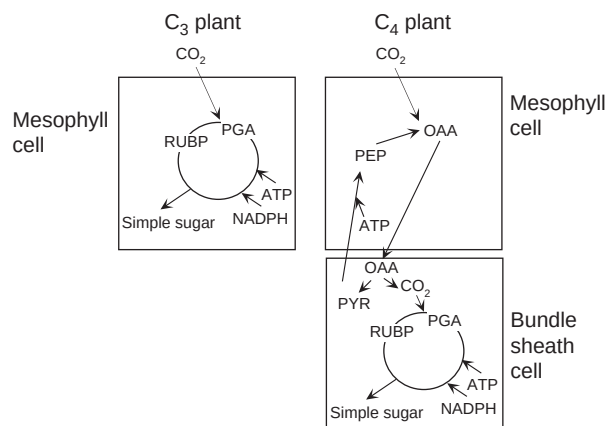


FIGURE 2 The leaf anatomy and biochemistry of a typical C_3 plant compared to a typical C_4 plant. The specialized structure and biochemistry allow C_4 plants to concentrate CO_2 inside the leaf and avoid losing carbon through photorespiration.

upper edge of the leaf surface. In contrast, plants utilizing the C_4 pathway have vascular bundles surrounded by a sheath of cells rich in chlorophyll. C_4 plants initially carboxylate the acceptor molecule phosphoenolpyruvate (PEP) using the enzyme PEP carboxylase to form oxaloacetate, which is then converted to one of two four-carbon organic acids, malic or aspartic acid (hence the term C_4 photosynthesis). These molecules are transported internally to the specialized bundle sheath cells. There, the plant enzymatically breaks down the malic and aspartic acid to release the fixed CO_2 . The CO_2 is then fixed again using the normal enzymes of C_3 photosynthesis to form three-carbon molecules. C_4 plants do not replace the C_3 cycle; the PEP in conjunction with the bundle sheath anatomy acts as a CO_2 pump to concentrate CO_2 at the site of Rubisco fixation in the bundle sheath cells. As a result, the concentration of CO_2 within the bundle sheath cells is much higher than that in the surrounding atmosphere and effectively eliminates photorespiration.

Under scarce water conditions, a third photosynthetic pathway has developed: crassulacean acid metabolism (CAM). These plants tend to be succulents and members of the plant families Cactaceae, Euphorbiaceae, and Crassulaceae. The CAM pathway is similar to the C_4 pathway in that CO_2 is first fixed into a four-carbon acid, malate, using the enzyme PEP carboxylase. The malate is then enzymatically cleaved, yielding a CO_2 molecule, which then participates in the traditional dark reactions of C_3 photosynthesis. Unlike C_4 plants, in which these two steps are physically separated in the mesophyll and bundle sheath cells, in CAM plants both steps occur in the mesophyll cells but at different times. The CAM plant opens its stomata at night, taking up CO_2 and fixing it to malate using PEP carboxylase. During the day, the plant closes its stomata and converts the malate into CO_2 , which is then fixed using the normal C_3 dark reactions. By only opening stomata at night when the temperature is lowest and relative humidity highest, CAM plants conserve water. Although CAM plants use water very efficiently, the pathway is energetically expensive and is not usually observed except in very dry environments.

6. GLOBAL CARBON CYCLING: PHOTOSYNTHETIC CONSIDERATIONS

The global cycling of carbon involves both biological and physical processes. Only the biological compo-

nents are discussed here. Plants and other photosynthetic organisms assimilate CO_2 and eventually convert the CO_2 to simple sugars. From simple sugars, plants synthesize a variety of compounds and store them in the form of plant tissue. When herbivores consume plant material, simple sugars are synthesized and transformed to other carbon-based compounds. Predators feeding on the herbivores redigest and again transform the carbon. At each step, some of the carbon returns to the atmosphere in the form of CO_2 as a by-product of respiration. The remainder is incorporated and temporarily stored in living biomass. Carbon contained in living organisms is eventually released through decomposition. The rate of release is highly variable and depends on both environmental and biological conditions. In tropical forests, most of the carbon in detritus is rapidly recycled. In drier, cooler regions considerable quantities of carbon are stored in the soil for long periods of time.

Similar cycling takes place in aquatic environments. Phytoplankton assimilate carbon dioxide that has diffused into the upper layers of water and dissolved carbonate and bicarbonate ions and convert it to organic sugars. The carbohydrate produced eventually passes through the aquatic food web. Significant portions of carbon, bound as carbonate in shells and exoskeletons of organisms, settle to the lake or ocean floor. This carbon is largely isolated from biotic activity and becomes incorporated into bottom sediments. It may then enter a long-term pool and may not recycle to the atmosphere again until it resurfaces as limestone through geological time.

The total global carbon pool involved in global carbon cycling is estimated to be approximately 55,000 Gt (1 Gt = 1 billion metric tons). Fossil fuels, the result of ancient photosynthesis, account for approximately 10,000 Gt. The ocean contains approximately 35,000 Gt, primarily as bicarbonate (HCO_3^-) and carbonate ions (CO_3^{2-}). Dead organic matter in the ocean contains 1650 Gt of carbon. Living biomass in the ocean, primarily phytoplankton, is 3 Gt. Terrestrial ecosystems contain approximately 1200 Gt as dead organic matter and 550 Gt as living matter. The atmosphere holds approximately 740 Gt of carbon.

Although global carbon cycling is a combination of both physical and biological processes, photosynthesis is obviously a central controlling factor. In order to understand the global carbon cycle, we must understand how photosynthetic reactions occur over broad temporal and spatial scales.

Human activities have severely altered the global carbon cycle during the past century, with unknown future consequences.

7. GLOBAL CLIMATE IMPLICATIONS

The flux of carbon among terrestrial, aquatic, and atmospheric pools is at least partially controlled by photosynthetic processes. The relative carbon pool sizes have major effects on global climate. The CO₂ flux among land, sea, and atmosphere has, since the industrial revolution, been disturbed by a rapid injection of carbon dioxide to the atmosphere from the burning of fossil fuels and from the clearing of forests. Both the fossil fuel and biomass carbon are relatively long-lived pools with long residence times. However, human activity has volatilized this carbon originally fixed by photosynthesis into the atmosphere. At the beginning of the industrial revolution, atmospheric CO₂ was approximately 260 ppmv. Since then, input of CO₂ has increased this concentration to approximately 372 ppmv in 2002.

Only approximately 40–50% of the total CO₂ injected into the atmosphere remains there. What happens to the rest is a subject of current research and not completely understood. Most likely, the ocean is a large sink for this excess carbon. The other potential major sink is the terrestrial biota. Terrestrial ecosystems, especially forests, do have the ability to store some carbon through the photosynthetic production of biomass. However, whether this storage is greater than the contribution of carbon through respiration is not clear in many terrestrial systems. A complete mechanistic understanding of the global carbon cycle has yet to be achieved, and this will be an area of active scientific endeavor for many years. However, there is very little disagreement that atmospheric CO₂ has increased dramatically during the past century and that if atmospheric CO₂ concentrations continue to increase at the present rate, it will exceed 400 ppmv by 2030. In view of our dependence on fossil fuels and the growing body of evidence linking atmospheric CO₂ and climate, understanding the fate of excess atmospheric carbon is one of the most pressing scientific research needs.

In addition to potentially changing the earth's climate system, rising atmospheric CO₂ concentrations directly impact the photosynthetic process and may increase the primary production of

photosynthetic organisms in some systems. Satellite imagery and remote-sensing information for the past two decades indicate as much as a 6% increase in total production of terrestrial systems, primarily in the tropics. Smaller scale manipulative experiments, such as free air CO₂ enrichment (FACE) studies, have yielded more complex results. Most FACE studies have observed an initial increase in productivity but then a stabilization in productivity over time. The mechanisms behind these observations are not clear, but it seems likely that some other limitation besides CO₂ is eventually expressed in most terrestrial systems. The processes and feedbacks of carbon cycling, of which photosynthesis is a major component, are very active areas of research.

8. CONCLUSION

The photosynthetic process is literally the most important energy transfer on Earth. The fixation of carbon by photosynthetic organisms provides the high-energy molecules needed to sustain nearly all organisms on Earth. Photosynthesis is a two-stage process: (i) harnessing of light energy to produce reducing power (NADPH) and energy (ATP) and (ii) use of the reducing energy to fix organic molecules. We usually associate photosynthesis with the reduction of atmospheric CO₂ to organic sugars. However, the energy generated during the light reactions of photosynthesis can be used to reduce other compounds (e.g., nitrogen and sulfur). Photosynthetic reactions have, in part, controlled the relative carbon pool sizes on Earth. Human activities, especially fossil fuel consumption and biomass burning, have led to an injection of carbon as CO₂ into the atmosphere. This rapid increase in carbon in the atmosphere has the potential to dramatically alter future Earth systems function and climate. A complete mechanistic understanding of global carbon cycling including photosynthetic processes is imperative.

SEE ALSO THE FOLLOWING ARTICLES

Carbon Sequestration, Terrestrial • Conversion of Energy: People and Animals • Earth's Energy Balance • Food Capture, Energy Costs of • Heterotrophic Energy Flows • Origin of Life and Energy • Photosynthesis, Artificial • Sun, Energy from • Thermoregulation

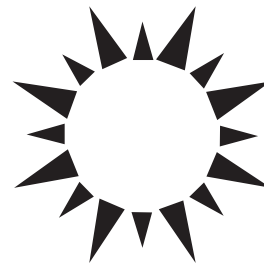
Further Reading

- Davis, G. R. (1990). Energy for planet Earth. *Sci. Am.* **263**, 20–27.
- Hall, D. O., and Rao, K. K. (1999). “Photosynthesis.” Cambridge Univ. Press, Cambridge, UK.
- Laisk, A., and Oja, V. (1998). “Dynamics of Leaf Photosynthesis: Rapid-Response Measurements and Their Interpretations.” CSIRO, Collingwood, Australia.
- Lawlor, D. W. (2001). “Photosynthesis.” Springer-Verlag, New York.
- Wigley, T. M. L., and Schimel D. S. (2000). “The Carbon Cycle.” Cambridge Univ. Press, Cambridge, UK.

Photosynthesis, Artificial

NATHAN S. LEWIS

California Institute of Technology
Pasadena, California, United States



1. Types of Artificial Photosynthetic Devices
2. Basic Processes of a Photoelectrochemical Energy Conversion Device
3. Photoelectrolysis Using Metal Oxides and Other Semiconductors
4. Approaches to Constructing Stable, Efficient Photoelectrochemical Cells

Glossary

- G0005 **anode** The electrode in an electrochemical cell that supplies electrons into the external circuit. This electrode results in removal of electrons from the electrolyte at the electrode/electrolyte interface.
- G0010 **cathode** The electrode in an electrochemical cell that injects electrons into the electrolyte.
- G0015 **electrochemical cell** A complete electrical circuit, consisting of two electrodes, the anode and the cathode, an electrolyte in contact with the electrodes, and a redox species that can carry electrical charge in the electrolyte solution.
- G0020 **electrolyte** A liquid containing an ionic salt that allows electrical conduction in the liquid.
- G0025 **photoelectrode** A material that absorbs incident light and produces current through an external circuit.
- G0030 **photoelectrolysis of water** A specific photoelectrochemical cell in which water is oxidized, forming oxygen gas, at the anode, and water is reduced, forming hydrogen gas, at the cathode.
- G0035 **redox couple** A pair of chemical species that is involved with donating and/or accepting electrons from electrodes in an electrochemical cell. The redox couple consists of a chemically reduced species that can donate electrons to an electrode, thereby forming the chemically oxidized form of the redox couple. The oxidized species of the redox couple can accept electrons from an electrode, forming the chemically reduced form of the redox couple.
- P0005 The global primary energy consumption from all sources in 1998 was approximately 400×10^{18} J, which equates to a mean energy consumption rate of 12.8×10^{12} W or 12.8 terawatts (TW). By comparison, the net solar power striking the surface of the earth is approximately 1.2×10^5 TW. A diurnally averaged

20 TW of power could be obtained if approximately 0.16% of the land on Earth were devoted to energy farms that converted solar energy at 10% net conversion efficiency. Biomass farms generally display net solar energy conversion efficiencies of 0.1–0.5%. In contrast, the theoretical maximum energy conversion efficiency for a simple light-absorbing solar energy conversion device is more than 30%, and photovoltaic cells that produce electricity from sunlight have demonstrated efficiencies in the field in excess of 20%. Numerous laboratories throughout the world are therefore building artificial photosynthetic systems with the hope of understanding, and ultimately improving on, the energy conversion properties of the natural photosynthetic process.

1. TYPES OF ARTIFICIAL PHOTOSYNTHETIC DEVICES

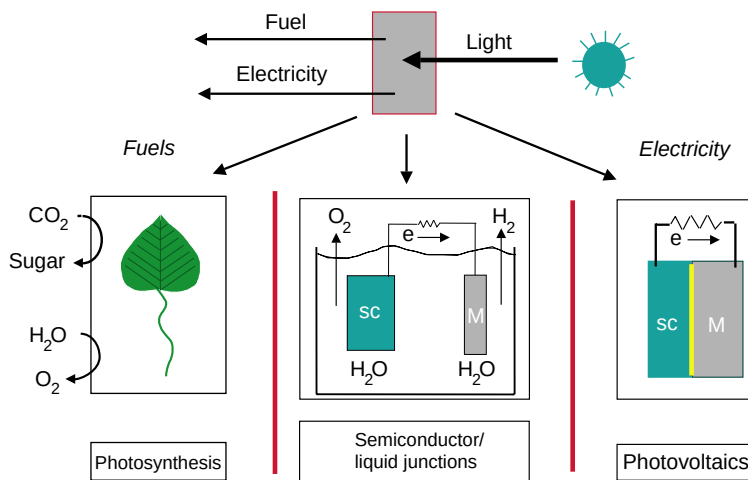
S0005

Several molecular systems have been constructed that mimic various aspects of photosynthesis. Two of these utilize molecular systems that are derived from natural photosynthesis but that incorporate chemically based modifications to produce artificial photosynthetic devices. These devices use artificial photosynthetic pigments to drive chemical reactions across lipid bilayers or use noble metal catalysts to change the function of the photosynthetic process to produce hydrogen and oxygen instead of sugars and oxygen. Neither of these systems are sufficiently robust to be operated for extended periods of time as energy conversion devices, but they have shown that it is possible to produce artificial photosynthetic assemblies that function well in a laboratory setting.

P0010

The third system, semiconductor photoelectrochemistry, utilizes a nonbiological approach to produce fuels in an entirely artificial photosynthetic system. Photoelectrochemical systems use sunlight to supply part or all of the energy required to generate the storable chemical fuels (Fig. 1). Photoelectrochemical

P0015



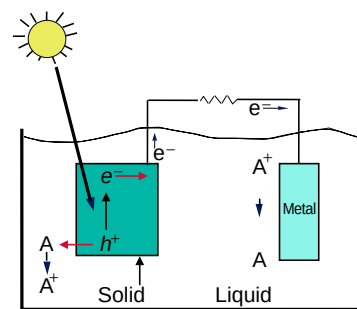
F005 **FIGURE 1** Known methods for transforming solar energy into chemical fuels or electrical energy. Photosynthesis stores approximately 3% of the incident solar energy in the form of chemical fuels for use by the plant. Photovoltaics produce electrical energy with up to 25% efficiency. Wet solar cells based on semiconductor/liquid junctions produce either storable chemical fuels or electrical energy, depending on their mode of operation. Efficiencies of these artificial photosynthetic systems can be greater than 15%.

systems can also be utilized to provide electrical energy. Photoelectrochemical energy conversion systems, although a relatively new technology, have demonstrated laboratory solar energy conversion efficiencies of >16% for electricity formation and >10% for fuel formation during the sunlight-assisted splitting of water into $\text{H}_2(\text{g})$ and $\text{O}_2(\text{g})$. In this respect, they represent a promising method for obtaining renewable, storable energy from the sun.

P0020 A photoelectrochemical cell is closely related to a battery, fuel cell, or other type of electrochemical system in that it is composed of two electrodes and an electrolyte. Unlike batteries or fuel cells, in which the energy is initially stored in the reactants and is then released through the electrical discharge circuit during the formation of chemical reaction products, photoelectrochemical cells utilize the input optical energy to drive electrochemical reactions (Fig. 1). To construct a photoelectrochemical cell, one of the metal electrodes of a conventional electrochemical cell is replaced by a semiconductor electrode. Light absorbed by the semiconductor electrode produces a current through the cell, which results in the production of chemical fuels and/or electrical energy by the device.

S0010 2. BASIC PROCESSES OF A PHOTOELECTROCHEMICAL ENERGY CONVERSION DEVICE

P0025 The basic processes that occur in such a system are well understood. The semiconductor electrode effi-



Light is converted to electrical and chemical energy

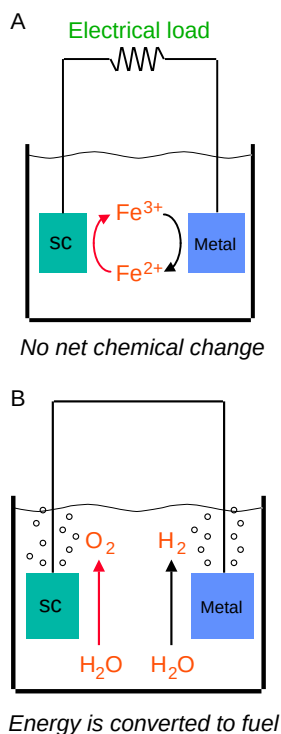
FIGURE 2 Basic operation of a photoelectrochemical cell. The semiconducting electrode absorbs energy from sunlight and produces a current through the electrochemical cell. The current produces electrical and/or chemical energy in the system.

ciently absorbs light, producing an excited electronic state. In this excited state, the electron and the electron vacancy (the “hole”) are both more energetic than they were in their respective ground states. The photo-excited electrons and holes are generally not tightly bound to an individual atom or set of atoms in the solid. These delocalized charge carriers readily move through the semiconductor and are therefore available to perform chemical reactions at the solid/liquid interface (Fig. 2). When the photogenerated electrons and holes drive chemical reactions at the electrodes in the cell, energy from sunlight is stored in the system.

If the chemical reaction at one electrode is exactly the opposite of the reaction at the other electrode, no net chemical change will occur in the electrolyte. The overall process is then conversion of light into

F0010

P0030



electricity. This mode of photoelectrochemical cell operation resembles that of a photovoltaic device and is designated as a regenerative photoelectrochemical cell. Alternatively, if the chemical reactions are different at each electrode, chemical fuels will be produced by the light-induced current flow. This type of system is designated as a fuel-forming photoelectrochemical cell. Both regenerative and fuel-forming photoelectrochemical cells have been demonstrated in the laboratory. A particularly appealing situation is depicted in Fig. 3, in which water is the liquid and current flow through the circuit simultaneously produces electricity, hydrogen gas, and oxygen gas.

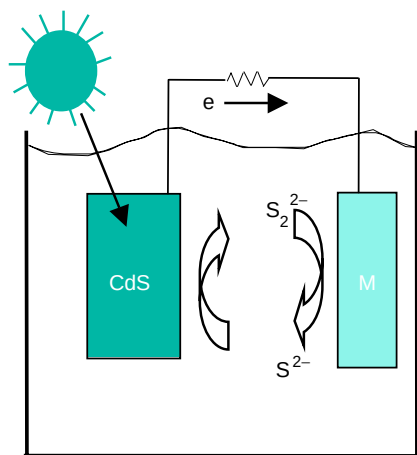
Closer consideration of a photoelectrochemical cell makes it clear that efficient light absorption alone is not sufficient to obtain efficient device operation. If the charged carriers created by optical excitation of the semiconductor had no preference regarding their direction of motion, equal numbers of excited electrons would travel in clockwise and counterclockwise directions through the circuit in the cell. The equal magnitudes, but opposite signs, of the currents would lead to a net cancellation of current in the wire, and no net energy output would be obtained from such a cell.

Charge separation, however, is very efficiently effected in a well-designed photoelectrochemical cell. The key to directing the charge carriers preferentially through the circuit lies in the formation of a strong electric field at the semiconductor/liquid interface (Fig. 4). This electric field either attracts or repels electrons from the interface, depending on the polarity of the field. Fortunately, maintaining the field does not require a constant energy input from an external source: This electric field occurs spontaneously whenever two phases with different chemical properties are brought into contact. In semiconductor/liquid junctions, the interfacial fields are generally so strong that any photogenerated electrons and holes are rapidly driven in opposite directions, and charge carrier recombination is thereby suppressed.

3. PHOTOELECTROLYSIS USING METAL OXIDES AND OTHER SEMICONDUCTORS

Certain materials, generally metal oxides, can sustain the unassisted photoelectrolysis of water into H_2 and O_2 (see Fig. 3). The energy contained in the photons that are absorbed by these materials can in many cases be efficiently converted into energy stored into

FIGURE 3 Two modes of operation for wet solar cells based on photoactive semiconductor materials. In both possibilities shown, the semiconductor (sc) absorbs the light and produces energetic electrons that can flow through the circuit. The differences between the cells result from the types of chemical reactions that occur in the liquid as a result of this electron flow. (A) A wet solar cell that produces electrical power from incident sunlight. The semiconductor produces excited electrons that flow through the circuit, doing work on an electrical load. This load could be a light bulb, battery, or other type of electrical machinery. The electrons are then injected into the solution at the metal electrode and produce a chemical reaction in the liquid. In this example, the injected electrons convert one form of iron (Fe^{3+} , with three ionized electrons) to another (Fe^{2+} , with only two ionized electrons). The Fe^{2+} then moves to the semiconductor electrode, carrying with it the injected electron. To complete the circuit, the Fe^{2+} then donates this electron to the semiconductor electrode. This regenerates the Fe^{3+} state in the solution. Thus, no net chemical change occurs in the liquid, but the light has produced a current through the electrical device in the circuit. (B) A wet solar cell that produces both electrical power and chemical fuels from incident sunlight. The processes of light harvesting and current flow through the circuit are very similar to those in A, but the chemical reactions are different at the two electrodes. In this example, the injected electrons react with the water to produce gaseous hydrogen (H_2), which bubbles out of the solution. To complete the electrical circuit, the semiconductor electrode must be provided with electrons. Here, they come from the water. Removal of electrons from the water produces oxygen gas (O_2), which also bubbles out of the solution. Thus, illumination of this cell with light produces both current flow and gaseous chemical fuels. The gases could be collected and stored for later use, with the stored energy being recovered either in a fuel cell or by burning the hydrogen and oxygen in an engine to regenerate the water.



this should be the case; rather, it is a limitation of the known materials systems in photoelectrochemistry.

4. APPROACHES TO CONSTRUCTING STABLE, EFFICIENT PHOTOELECTROCHEMICAL CELLS

4.1 Regenerative Cells in Aqueous Solution

Several approaches are being pursued to circumvent the problems described previously with the direct, efficient photoelectrochemical splitting of water. An alternative approach to minimize aqueous-based corrosion of semiconductor electrodes has been to search for water-soluble molecules that exchange electrons with the photoelectrode much more readily than does water. Such molecules, if present in the liquid in sufficient concentration, could prevent electrode corrosion by reacting with the photogenerated holes in the electrode before the holes can participate in the corrosion process. If the oxidized form of these corrosion-inhibiting reagents also provides a useful fuel, the system is even more desirable.

This approach has been successfully implemented for the II–IV semiconductors, such as CdS, CdSe, and CdTe. These materials can be stabilized in aqueous solutions through addition of sulfide to the solution to provide another source of reduced sulfide to compete with electrode corrosion (Fig. 4). In this manner, a stable regenerative photoelectrochemical cell can be constructed from an otherwise unstable semiconductor electrode. Long-term stability for several years has been demonstrated for such systems. The use of the stabilizing agent eliminates the possibility of production of desirable chemical fuels from the photoelectrochemical process but in return provides a stable photoelectrochemical cell for the conversion of light into electrical power. For example, stable, 16% efficient cells have been obtained using gallium arsenide in contact with an aqueous medium containing high concentrations of selenide ions.

CuInSe₂ photoanodes have yielded >10% efficiency for the production of I₂ from aqueous iodide ions. The production of I₂ is more desirable than the formation of Se (as in the GaAs system) because the I₂ could provide a useful fuel to charge an *in situ* storage battery. If properly coupled with a metal electrode, this CuInSe₂ cell could produce H₂ and I₂

FIGURE 4 Schematic of a wet semiconductor/liquid-based solar cell with a stabilizing agent in the solution. The light is absorbed by the semiconductor and produces excited electrons. Although CdS is shown, the semiconducting material could be silicon, gallium arsenide, titanium dioxide, or other suitable samples. The excited electrons are directed into the wire by the electric field because the negatively charged electrons are repelled by the negative charges that reside on the liquid side of the semiconductor/liquid boundary. The electrons then flow clockwise through the circuit, performing work through a load. They then reach the metal electrode and are forced to undergo a chemical reaction with the liquid in order to continue the circuit. The chemical produced from this reaction then moves through the liquid and reacts with the semiconductor electrode. This restores the electron to its original position and completes the current loop. The stabilizer is needed to accept and donate electrons at the two electrode/electrolyte interfaces to prevent corrosion or passivation of the two active electrodes in the cell.

the chemical fuels produced by the photoelectrolysis system. However, these materials have optical absorption thresholds (band gaps) that are too large to permit effective absorption of the visible and infrared portions of the solar spectrum, thereby wasting a large fraction of the incident solar energy and yielding overall energy conversion efficiencies of <1%. Modification of metal oxide materials has been claimed to be much more promising, with a modified TiO₂ photoanode yielding 8% efficiency in the photoassisted splitting of water into H₂ and O₂.

Metal chalcogenides have been investigated as photoanodes because the lattice S, Se, or Te atoms are weaker oxidants than lattice oxygen atoms and thus would enable more efficient utilization of the incident solar energy. Photoelectrochemical studies of metal chalcogenide electrodes, however, have revealed materials that are better absorbers than TiO₂ and therefore can provide more desirable semiconductor electrodes in photoelectrolysis systems. However, these materials are unstable in contact with water. There is no fundamental thermodynamic reason that

S0020

S0025

P0055

P0060

P0065

from aqueous HI. The stored chemical energy in the H_2 and I_2 would then be recovered at a later time, much as would be done if H_2 and O_2 gases had been obtained from a water-splitting process. Similar results for I_2 production have been obtained for $MoSe_2$ electrodes and for *in situ* electrical storage using the alloy material $CdSe_{0.65}Te_{0.35}$. Si electrodes have been covered with small metal islands to produce Br_2 from aqueous HBr with high efficiency, and $>10\%$ solar energy conversion efficiencies have been demonstrated for the formation of $H_2(g)$ from water at p-type InP electrodes. The key to making these fuel-forming systems more practical is the efficient coupling to the other, nonsemiconducting electrode: If this can be done without significant additional electrical energy input, fuel production could be sustained in a very efficient manner.

ductor and the metal electrodes in the cell. Rapid reaction with the semiconductor is required to compete effectively with corrosion reactions that will arise from the inevitable presence of trace quantities of water in the cell, whereas rapid electrochemical reaction with the metal electrode will minimize resistive losses from the charge transfer step. Nonaqueous semiconductor/liquid junctions are almost universally restricted to operating as regenerative photoelectrochemical cells because the reduced and/or oxidized forms of the stabilizers are rarely useful fuels. The stabilizing reagents are therefore most effectively utilized when they are added to the cell only once to facilitate the sustained stable production of electrical energy from an otherwise unstable semiconductor electrode.

An advantage of nonaqueous photoelectrochemical cells is that they allow utilization of many semiconductors that are not stable in any known aqueous electrolyte. An example is provided by the properties of Si photoanodes, which are readily stabilized for months when the solvent is methanol or ethanol in the presence of organometallic stabilizing agents (Fig. 5). Another advantage of using nonaqueous solvents is the ability to use a variety of stabilizing agents and to vary their chemical properties widely while still obtaining stable

P0080

S0030 4.2 Regenerative Cells in Nonaqueous Solvents

P0070 For some materials, such as CdS, several stabilizer/electrolyte combinations have been identified that are stable on the months/years timescale, whereas for others, such as Si photoanodes, no reagent has been discovered that provides stable operation even on the timescale of minutes in aqueous solution. A general approach to circumvent the instability of electrode materials such as Si and GaAs in aqueous solutions is simply to eliminate the corrosive liquid, water, from the environment of the photoelectrochemical cell. If the water in the cell is replaced by alcohol or another organic liquid, stable photoelectrode operation can readily be obtained, provided that the organic solvent also does not effect corrosion of the solid. Using this approach, stable photoelectrochemical cells have been obtained with Si, GaAs, InP, and other semiconductors whose optical properties are well matched to the solar spectrum.

P0075 In a nonaqueous photoelectrochemical cell, photogenerated charge carriers in the semiconductor are not scavenged by reaction with the solvent but are transferred to a chemical reagent that has been deliberately added to the electrolyte solution. This chemical stabilizing agent accepts an electron (is reduced) at one electrode and donates that electron (becomes oxidized) at the other electrode. The stabilizing agent carries the electrical charge between the plates of the electrochemical cell but undergoes no net overall chemical change during extended cell operation. Stabilizers should ideally be species that rapidly exchange electrons with both the semicon-

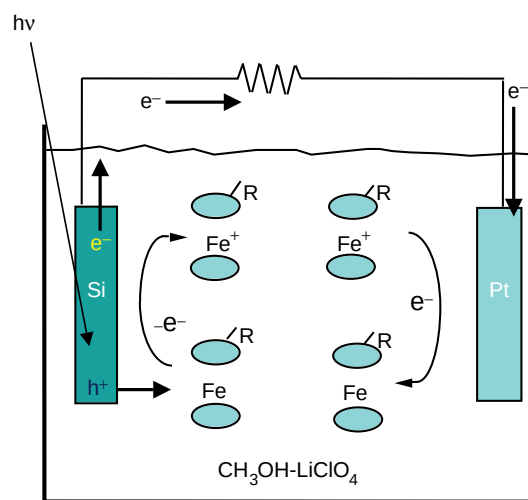


FIGURE 5 Stabilization of semiconductor electrodes in nonaqueous solvents. Although unstable in aqueous solutions, many semiconductors can be stabilized in nonaqueous electrolytes. In this example, the Si electrode is in contact with a nonaqueous liquid that contains a high-boiling solvent (propylene carbonate), salt to provide electrical conductivity (lithium perchlorate), and redox ions to carry the electrical charges through the liquid (ferrocene, which is an organometallic complex of Fe^{2+} that is soluble in organic solvents). The cell is 14% efficient in converting incident solar energy into electrical energy.

F0025

electrical currents from the photoelectrochemical device. This can be extremely useful in optimizing the energy conversion efficiency parameters of a photoelectrochemical system. For example, in the water splitting systems discussed previously, significant energy is wasted when there is a large energy difference between the levels of electrons or holes in the photoelectrode and the oxidation/reduction potentials for the water splitting reaction. Such energy differences can be minimized, for example, using Si/alcohol contacts because the energy level of the stabilizer can be changed until it closely matches the desired energy level of the semiconductor. This allows optimization of the cell efficiency through simultaneous chemical design of the electrolyte, photoelectrode material, and stabilizing agent.

P0085 These nonaqueous photoelectrochemical devices can yield relatively high solar energy conversion efficiencies: 12–14% efficiency has been obtained in unoptimized photoelectrochemical cells compared to the 15–20% efficiency that is obtained for production-line solid-state Si and GaAs photovoltaics. The nonaqueous photoelectrochemical cells also offer an opportunity to readily change the chemistry at the solid/liquid interface and to investigate the effect of these changes on the electron yield of the system. Although much remains to be learned about the principles that govern the behavior of these cells, many of them can properly be denoted as technical successes because they are efficient and stable in sunlight.

S0035 4.3 Dye Sensitization of Stable Electrode Materials

P0090 Another approach to artificial photosynthesis using semiconductor photoelectrochemistry is to extend the optical response of photoelectrode materials that are known to be stable in water, such as TiO₂. When a highly colored dye is coated on the surface of a semiconductor, the absorption of a white (i.e., high band gap) material such as TiO₂ can be extended into the visible region of the spectrum. This type of dye sensitization is analogous to the chemistry that forms the basis of the modern color photographic process, with the solid being semiconducting AgBr and with various dyes providing the color response of the AgBr film emulsion.

P0095 Most dye-coated, single-crystal semiconductor electrodes exhibit solar energy conversion efficiencies of less than 0.1%. Even with the most highly colored molecules (i.e., dye molecules having very

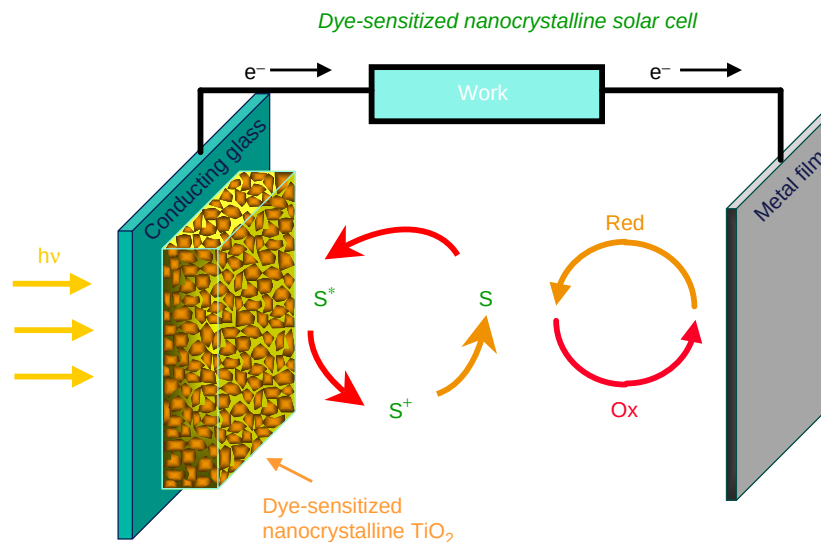
large optical absorption coefficients), several layers of dye are required to absorb all the incident light. However, only the dye layer closest to the electrode surface is generally effective in injecting an excited electron to the electrode. Since only a monolayer of dye is effectively useful for light absorption, most of the photons incident onto the cell are either transmitted through the device or absorbed by dye molecules that do not contribute to photocurrent flow.

To deal with this issue, very rough, high-surface-area, dye-coated titanium dioxide electrodes are used. In this system, the TiO₂ consists of crystallites with a diameter of approximately 10 nm. A sintered film of these nanocrystals, although relatively transparent, possesses such a high surface area that only a monolayer of adsorbed dye molecules can absorb >90% of the photons incident onto the sample. The excited state of the adsorbed dye is sufficiently reducing that it will donate an electron to the conduction band of the TiO₂. The presence of the TiO₂, or similar semiconducting material, is crucial because the TiO₂ accepts the photogenerated electrons from the excited state of the dye and directs the injected electron toward the electrical contact, preventing recombination with the oxidized form of the dye or with the species in the electrolyte solution. The dyes investigated to date, including Ru(polypyridyl) complexes and chlorophylls, have extended the absorption range of TiO₂ into the visible region of the spectrum and have provided much better overlap of the electrode response with that of the solar spectrum. Furthermore, the excited states of these adsorbed dye molecules have yielded very effective charge separation, producing photocurrent in the TiO₂ with near unity quantum yield. The efficiency of the system is thus not limited by the charge collection and electron injection process but by the voltage that can be obtained from the specific system of concern.

This system resembles the molecular photosynthetic process in that the photochemical reactions in the cell are completely dictated by the reactivity of the adsorbed dye molecules. The solid induces the charge separation and dictates the directionality of current flow, but the adsorbed dye molecules determine the course of the remaining steps in the energy conversion cycle. After injection of an electron into the solid, the oxidized dye molecule receives an electron from a soluble electron donor, regenerating the original form of the dye. The electrical circuit is then completed by injection of the electron into the solution at the other electrode in the system (Fig. 6).

P0100

P0105



F0030 **FIGURE 6** A wet solar cell made from a dye-sensitized semiconductor. A colored dye is used to allow normally transparent semiconductors to absorb sunlight in the visible region of the electromagnetic spectrum. The high surface area of the mesoporous solid allows nearly complete absorption of the incident light by a monolayer of dye adsorbed onto the surface of the semiconducting solid. The excited electrons produced from absorption by the dye are injected into the semiconductor and then make their way through the circuit as in a conventional wet cell. The dye is a ruthenium-based metal complex, and the support is a high-surface-area TiO_2 material, related to the inexpensive pigment commonly used in white paint. Solar energy conversion efficiencies as high as 10% have been reported in test devices of this type of potentially inexpensive solid/liquid junction.

P0110 Some of the Ru(polypyridyl) dyes investigated have oxidation potentials that are very close to those required to form $\text{O}_2(\text{g})$ from water. Thus, the problems encountered in the photoelectrochemistry of metal oxide electrodes, involving severe energy loss between the level of the photogenerated holes and potential for the oxidation of water to $\text{O}_2(\text{g})$, can be avoided, in principle, in the dye-sensitized system. The electronic energy levels of the adsorbed dye molecules are simply tuned chemically until they are close to that of the desired chemical reaction in the cell. Unfortunately, oxidation of water to oxygen gas is a multielectron process and is generally very difficult to accomplish at a rapid rate. To quickly regenerate the original dye molecule for reuse in another photochemical cycle, another molecule, called a sensitizer, must be added to the dye-sensitized photoelectrochemical cell. The sensitizer donates an electron to the oxidized form of the dye, regenerating the original form of the dye and producing the oxidized sensitizer. The oxidized sensitizer is then reduced at the second electrode, yielding a regenerative photoelectrochemical cell for the entire photoelectrochemical cycle.

P0115 The TiO_2 dye-sensitized photoelectrochemical cell has been modified and improved by using nonaque-

ous solvents. Nonaqueous solvents allow greater stability of the system and also a wider choice of sensitizers. This system is essentially an extension of the strategy for producing stable photoelectrochemical cells involving the use of redox stabilizers in nonaqueous electrolytes, but in which relatively expensive single-crystal semiconductors have been replaced by less expensive, dye-sensitized nanocrystalline photoanode materials. Using this approach, solar energy conversion efficiencies of up to 10% have been reported. Relative to conventional photovoltaic cells, it has been claimed that this type of photoelectrochemical system has cost advantages, including the use of inexpensive raw materials, the lack of difficult fabrication steps or expensive crystal purification processes, and the ability to implement the design simply over large areas. The system indeed appears promising for electrical energy generation, provided that efficiencies can be improved and that light-induced degradation of the dye and the other constituents in the cell can be kept to levels required for long-term, maintenance-free operation of the system.

As can be seen, there are many possibilities for improving these photoelectrochemical cells, and no one strategy has emerged as the clear choice. We know a great deal about the fundamental operating

P0120

principles of each component of these artificial photosynthetic systems. Advances have also been made in practical efforts to improve their energy conversion efficiencies and to increase their production rates of chemically desirable fuels. In some cases, the systems show promise in the laboratory as highly efficient energy conversion devices. Even for these systems, choices must still be made to obtain the correct combination of materials that might comprise a complete artificial photosynthetic system. Siting considerations indicate that at 10% overall conversion efficiency, a relatively small fraction of the suitable land area available globally would be required to produce significant amounts of power relative to current global fossil energy consumption. Artificial photosynthetic systems, however, are in their infancy, and the resultant energy costs that might be realized from a fully developed artificial photosynthetic technology are difficult to estimate. Possible advantages of the approach relative to photovoltaics for production of electricity are the ability to directly form fuels from an inherently intermittent energy source as well as the ability to use inexpensive raw materials as the light absorbers without suffering large decreases in energy conversion efficiency of the resulting photoelectrochemical energy conversion devices. The environmental impacts of these systems would be relatively benign provided that inert raw materials, such as Si and TiO_2 , were used as the basic photoelectrode materials. The actual decision to construct a commercial system will also require input regarding a number of market-related factors, including the price of alternative energy sources, the value of the fuels produced by the photoelectrochemical cells, and the cost in both dollars and energy of fabrication and maintenance of the system. There are similar constraints for all alternative energy sources, but the promise and demonstrated efficiency of these devices is too great to rule out these interesting artificial photosynthetic devices as energy options for the future.

SEE ALSO THE FOLLOWING ARTICLES

Photosynthesis and Autotrophic Energy Flows • Photovoltaic Materials, Physics of • Solar Fuels and Materials

Further Reading

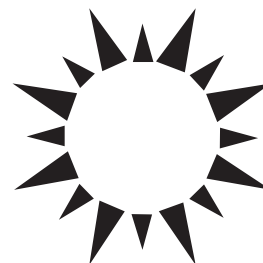
- Bard, A. J., and Fox, M. A. (1995). Artificial photosynthesis: Solar splitting of water to hydrogen and oxygen. *Accounts Chem. Res.* **28**, 141–145.
- Finklea, H. O. (1988). "Semiconductor Electrodes." Elsevier, New York.
- Fujishima, A., and Honda, K. (1972). Electrochemical photolysis of water at a semiconductor electrode. *Nature* **238**, 37–38.
- Gerischer, H. (1970). Semiconductor photoelectrochemistry. In "Physical Chemistry: An Advanced Treatise" (H. Eyring, D. Henderson, and W. Yost, Eds.), pp. 463–542. Academic Press, New York.
- Hodes, G., Fonash, S. J., Heller, A., and Miller, B. (1985). Photoelectrochemical cells based on polycrystalline semiconductors. *Adv. Electrochemistry Electrochem. Eng.* **13**, 113–158.
- Koval, C. A., and Howard, J. N. (1992). Electron transfer at semiconductor electrode–liquid electrolyte interfaces. *Chem. Rev.* **92**, 411–433.
- Nozik, A. J. (1978). Photoelectrochemistry: applications to solar energy conversion. *Annu. Rev. Phys. Chem.* **29**, 189–222.
- O'Regan, B., and Grätzel, M. (1991). A low-cost, high-efficiency solar cell based on dye-sensitized colloidal TiO_2 films. *Nature* **353**, 737–740.
- Parkinson, B. A. (1983). An overview of the progress in photoelectrochemical energy conversion. *J. Chem. Education* **60**, 338–340.
- Rajeshwar, K. (1985). Materials aspects of photoelectrochemical energy conversion. *J. Appl. Electrochem.* **15**, 1–22.
- Tan, M. X., Laibinis, P. E., Nguyen, S. T., Kesselman, J. M., Stanton, C. E., and Lewis, N. S. (1994). Principles and applications of semiconductor photoelectrochemistry. *Prog. Inorganic Chem.* **41**, 21–144.
- Tributsch, H. (1986). Photoelectrolysis and photoelectrochemical catalysis. In "Modern Aspects of Electrochemistry" (J. O. M. Bockris and B. E. Conway, Eds.), pp. 303–355. Plenum, New York.
- Wrighton, M. S. (1979). Photoelectrochemical conversion of optical energy to electricity and fuels. *Accounts Chem. Res.* **12**, 303–310.

Photovoltaic Conversion: Space Applications

PETER A. ILES

Arcadia

California, United States



-
1. Special Features of Space Photovoltaics
 2. Evolution of Space Solar Cells
 3. Progress in Photovoltaics
 4. Future Trends in Space Photovoltaics
 5. Conclusion

Glossary

air mass zero (AMO) Location outside Earth's atmosphere.

antireflecting coating Dielectric layers deposited on solar cells to reduce reflection of incident sunlight.

back surface field (BSF) Electric field at the back surface of a solar cell.

back surface reflector (BSR) Mirror formed at the back surface of solar cells to increase reflection of unabsorbed sunlight from the back surface.

beginning of life (BOL) Period immediately after spacecraft is launched.

conversion efficiency Fraction of the power output of a solar cell to the incident power from the sun, expressed as a percentage.

coverglass Transparent cover bonded to the front surface of solar cells.

end of life (EOL) Period after solar cells have withstood defined operating conditions in space, mostly exposure to radiation.

epitaxial layers Layers grown to replicate the crystal lattice properties of a substrate material.

metal organic chemical vapor deposition (MOCVD) Deposition of epitaxial layers using gaseous elements, obtained mainly from compounds consisting of organic chemical groups (e.g., methyl and CH_3) and a metal.

organometallic vapor phase epitaxy (OMVPE) Similar to MOCVD.

PN junction Voltage barrier formed in a semiconductor, providing the conversion of optical (photonic) energy into electrical energy.

P/N solar cell The illuminated surface is a P-type emitter layer formed over an N-type base. N/P solar cells have opposite layer polarity.

substrate The starting material for forming solar cells; also refers to the material to which solar cells are bonded to form panels (or arrays).

Photovoltaic arrays on satellites and spacecraft provide electric power for on-board equipment and communication with control centers on Earth. Before 1957, when the first satellite, Sputnik I, was launched, there was much debate about the best way to generate electric power on satellites. Most conventional fuels do not function well in the vacuum of space, and even if usable, they cannot operate for extended periods without exceeding the weight limits set by the propulsion rockets and posing serious safety concerns. Sunlight is the obvious power source in space, removing the need to launch fuel with the spacecraft. In 1954, Bell Telephone Laboratories (BTL) announced that it had made a significant increase in the output from silicon solar cells. Four years later, commercial versions of BTL cells were successfully used on the second U.S. satellite Vanguard I to provide power for a radio-tracking signal that lasted for several years. Following this successful demonstration, space photovoltaic (PV) power was rapidly accepted and used widely. For more than 40 years, arrays of solar cells have delivered power flawlessly for long periods. PV power is one of the major reasons for the successful performance of thousands of satellite. Satellite applications have become familiar, including worldwide telecommunications, sensing of weather and other Earth surface geographic conditions, and navigation using the constellation of global positioning satellites (GPS). Satellites have also performed many scientific missions, carrying astronomical telescopes and exploring around or on the moon.

and Mars and diverse locations toward or away from the sun. Military satellites have collected large amounts of data for surveillance or information gathering. Space PV continues to be an active niche market, supplied by organizations throughout the world. The high visibility of space programs has publicized the many advantages of PV-generated power.

1. SPECIAL FEATURES OF SPACE PHOTOVOLTAICS

Space photovoltaic (PV) differs from terrestrial PV in several ways.

1.1 Components of a Space Solar Power System

Space solar power systems comprise the following components:

- *Solar cells:* The most important components are the solar cells, which are connected together in series-parallel circuits to supply the required voltages and currents.
- *Coverglasses:* Thin transparent covers of glass or quartz are adhesively bonded to the cells to reduce their temperature by increasing the reradiation of infrared wavelengths into space. Covers also protect the cell from damage caused by low-energy protons and have an optical coating to reject ultraviolet (UV) wavelengths to prevent UV darkening of the adhesive.
- *Metal interconnectors:* Thin metal foils, 100–150 μm thick, are soldered or welded to electrically connect the contact pads on the cells. Placing all the contact pads on the back surface of the cells facilitates automated bonding. The metal-to-cell bonds must be robust for high reliability.
- *Array substrates:* Interconnected cells are bonded to lightweight substrates to form a panel or array. Aluminum honeycomb and graphite-epoxy composites are used most. Kapton has been used on fold-out arrays to reduce weight.
- *Deployment and tracking mechanisms:* In some cases, cells are attached to the outer surface of the satellite (body mounted), but they are usually mounted on wings or paddles. Wings are folded during launch and deployed by releasing spring clamps, or by folding out or rolling out, with means to lock the array in position. Sun tracking mechanisms keep the solar arrays facing the sun.
- *Power storage:* Chemical batteries store PV-generated energy and stabilize the electric power fed to the satellite circuitry. Many charge–discharge cycles can occur, and to operate in the vacuum, custom-manufactured sealed batteries are used with nickel–cadmium or lithium electrodes.
- *Electric power distribution:* Autonomous power conditioning equipment controls the battery charge conditions and allocates power to the circuits in the satellite.

1.2 Variety of Locations

Locations where spacecraft must operate are classified into four broad categories:

- *Near-Earth:* In the vicinity of the earth, there are five types of satellite orbits:
 1. Low earth orbit (LEO), altitude 150–1000 miles.
 2. Medium earth orbit (MEO), altitude 1000–22,000 miles.
 3. Geosynchronous earth orbit (GEO), altitude approximately 23,000 miles. Worldwide communications requires 3 satellites in GEO, 15–20 satellites in MEO, and 50 or more satellites in LEO.
 4. Orbital transfer missions (e.g., LEO to GEO).
 5. Cislunar to near the moon, 23,000–250,000 miles distant.
- *Near-Sun:* These include missions toward Venus, Mercury, or even closer to the sun.
- *Far-Sun:* Missions beyond the earth's orbit, including those to Mars, the asteroid belt, Jupiter, and Saturn.
- *Landers:* These are bases on the moon or Mars or for landing on comets or asteroids. Manned bases require a large amount of power.

1.3 The Space Environment

Space contains only minute amounts of gases. In this ultrahigh vacuum, heat cannot be removed from the spacecraft by conduction or convection but only by radiation. The power received from the sun and the solar array temperatures at any location depend on the distance from the sun. In the vicinity of the earth, outside the gaseous atmosphere, the incident solar output as a function of wavelength is shown as solar extraterrestrial in Fig. 1. This solar spectrum is called air mass zero (AMO), and the power density is approximately 1365 watts per square meter (Wm^{-2}).

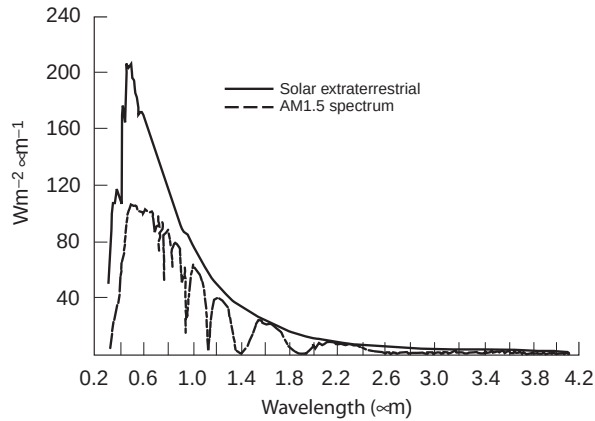


FIGURE 1 Solar spectrum incident on the earth. Solar extraterrestrial, outside gaseous atmosphere; AM1.5 spectrum, at the earth's surface.

Figure 1 also shows the spectrum received at the earth's surface (AM1.5), with the losses in the atmosphere reducing the power density to 800–1000 Wm^{-2} . All solar cell conversion efficiencies given in this article are AMO efficiencies, given by the percentage ratio of the electric power output of the cell, held at a test temperature of 30°C, to the incident power, 1365 Wm^{-2} , from a light source that simulates the AMO spectrum.

1.3.1 Charged Particles

Exposure to charged atomic particles emitted from the sun is not a problem, except during short periods of solar flares or when satellites move through the Van Allen Belts (VABs), in which the particles have been trapped by the earth's magnetic field. A series of satellites measured the energies and concentrations of charged particles in the belts. The outer VAB, at an altitude of approximately 14,000 miles, is more intense than the inner VAB, which has an altitude of approximately 3000 miles. The relative exposures to radiation in LEO, MEO, and GEO are low, high, and moderate, respectively. To estimate the power lost by solar cells from particle bombardment during a mission, the accumulated dosage for the specific mission orbit is calculated and used in ground radiation tests of cells to estimate the end of life output.

1.3.2 Solar Storms

It is recognized that very intense emissions from active regions on the sun can cause serious problems for both terrestrial and satellite power systems. The emissions include X-rays, bubbles of plasma ejected from the sun's corona (coronal mass ejection), and

high-energy charged particles, mostly protons and electrons. The plasma generates shock waves that accelerate atomic particles, including those trapped in the VAB, and it also depresses the earth's magnetic field to altitudes at which it affects satellites in GEO orbits. The charged particles cause electrostatic charging problems in solar arrays and in semiconductor circuitry. Failures in satellite attitude controls, especially if they are based on magnetometers, lead to loss of power when the solar arrays are not facing the sun and to reduced communication with the satellite antennas. The large increase in charged particles causes significant power loss from radiation damage in the solar panels.

These solar storm activities have occurred mostly during solar maxima, but problems are observed during quieter periods. Following disruption of several communications satellites, with worldwide impact, additional satellites have been deployed to monitor the sun's activity to provide early warning, allowing some preventive measures to be taken.

1.3.3 Other Orbital Effects

Satellites in low LEOs experience atmospheric drag, causing the satellites to move to lower orbits and eventually fall to the earth. To reduce the amount of booster rocket fuel needed to offset drag effects, the area of the solar arrays must be as small as possible.

For orbits below GEO, satellites pass into the shadow of the earth for part of the orbit. For LEO, shadowing for 30 min during each 90-min orbit causes temperature swings from 60 to -150°C . In 1 year, satellites in LEO can experience 9000 of these temperature cycles. For reliable operation during these cycles, qualified robust procedures must be used to deposit and bond the metallic contacts.

Solar arrays may experience impact damage from many thousands of pieces of space debris left by earlier satellites. The debris must be monitored, shielded, or avoided, especially for manned satellites.

1.4 Remote Operation with High Reliability

The PV power system must be highly reliable because its performance is essential to the success of the satellite mission. PV arrays usually operate unattended, although there have been a few cases of repair or replacement. The solar cells and the rest of the array are manufactured to comprehensive specifications, controlled jointly by the manufacturers and the users. The successful record of space PV arrays is

attributed to this tight process control and detailed documentation.

2. EVOLUTION OF SPACE SOLAR CELLS

Solar cells, which convert sunlight to electrical energy, are the most important component in space PV power systems. High beginning-of-life (BOL) efficiencies are needed because a limited area is available for mounting cells and arrays on satellites. High end-of-life (EOL) efficiencies and very high reliability when operating in space are also required.

2.1 Silicon Space Solar Cells

Silicon (Si) has a band gap (1.15 eV) slightly below the optimum value (1.35 eV) that gives highest PV efficiency. For many years, Si cells had the highest efficiency because of the high-quality, single-crystalline Si developed by the semiconductor industry. The efficiency of the P/N Si space cells derived from the Bell Telephone Laboratory cells was improved from 7 to 10%, and they were used until the early 1960s. After the discovery of the VABs, ground-based particle accelerator tests showed that N/P Si cells had superior resistance to radiation damage. N/P Si cells were quickly adopted and were steadily improved to become the workhorse solar cells used in space.

2.1.1 Choice of Silicon

After extensive testing, simple crystals of Si grown by the Czochralski method (Cz-Si) were found to give cells with the best combined BOL and EOL performance. Two Si resistivity levels were used most, 2 and 10 ohm-cm, containing boron (the P-type impurity), at concentrations of approximately 150 and 30 parts per billion, respectively.

2.1.2 Silicon Cell Improvements

Improved performance was obtained from cells that were manufactured under space-qualified manufacturing methods and that were also compatible with the design and assembly of space solar arrays. One improvement formed shallower PN junctions with reduced concentrations of impurity (phosphorus) in the N-layer. Back surface fields (BSFs) were incorporated to reduce loss of charge carriers by recombination at the back surface. Reflection of incident sunlight was reduced by depositing multilayer anti-

reflection layers, or by texturing the front surface by preferential etching, to form many small random pyramids. Cell temperatures in space were reduced by increasing the reflectivity of the back surface (BSR). Optimized gridline contacts were evaporated on the front surface, and the adhesion of the metal-to-Si contacts was increased.

For some special array designs, or to facilitate automated bonding procedures, pad contacts to the N and P regions were both located on the back surface of the cell. If the back contact was also formed in a grid pattern, the cells were bifacial, generating power when illuminated from the front and back. Bifacial cells can generate additional power in space, converting albedo reflection from clouds or the oceans. Thin Si cells (down to 70 μm) were manufactured. Table I shows production median BOL and EOL efficiencies for the three N/P Si cell designs used most until the 1980s and that are still used when they meet all mission goals. The EOL conditions specified in Table I are also used for EOL efficiencies mentioned in the following sections for other types of cells.

2.1.3 Advanced Silicon Cells

Research centers, principally the University of New South Wales and Stanford University, made significant increases in the efficiency of Si cells (to 19–20% AMO) by using high-quality float-zone refined Si (Fz-Si), by forming regular patterns of pyramids on the front surface, and by “passivating” the front and back surfaces to reduce recombination of charge carriers. Very clean processing conditions were maintained throughout processing to preserve the high quality of the Si. When advanced cells were adapted for space use, improved Cz-Si was used, with slightly lower BOL efficiency (17–18%). EOL efficiency was 12%,

TABLE I

Performance of Most-Used Silicon Space Cells^a

Cell structure	BOL efficiency (%) ^b	EOL efficiency (%) ^c
2 ohm-cm, BSR	13.5	10.0
10 ohm-cm, BSF, BSR	14.8	9.6
10 ohm-cm, BSR	12.5	10.1

^a Abbreviations used: BOL, beginning of life; EOL, end of life; BSR, back surface reflector; BSF, back surface field; AMO, air mass zero.

^b AMO illumination; cell temperature, 30°C.

^c AMO illumination; cell temperature, 30°C; after exposure to IE15 1 MeV electrons/cm².

higher than that of standard Si cells, but the large excess BOL efficiency resulted in problems in the power conditioning equipment. Figure 2 shows the structure of an advanced space Si cell. These cells have been used in applications when radiation exposures were moderate, but they were largely superseded by III–V cells.

2.2 III–V Space Cells

Compounds with elements in group III and group V in the periodic table are semiconductors, like the group IV elements Si and Ge, and have a wide range of properties.

2.2.1 Gallium Arsenide Solar Cells

Gallium arsenide (GaAs) has a band gap of 1.42 eV, close to the value giving peak solar cell efficiency. High-efficiency GaAs cells had been demonstrated, but the space cell community made significant improvements in forming large-area, high-efficiency GaAs cells. The important advance was the ability to grow the GaAs layers by organometallic vapor phase epitaxy (OMVPE) in large production-scale reactors that accommodated an area of 1500 cm² of GaAs substrates per run and gave very uniform layers over large areas. In a further advance, the GaAs substrates were replaced by Ge substrates, which have atomic lattice spacing very similar to that of GaAs, leading to growth of high-quality epitaxial GaAs layers. Ge substrates cost less and are stronger mechanically,

allowing fabrication of GaAs/Ge large-area, thin GaAs cells. GaAs/Ge cells with an area of more than 36 cm² with median BOL efficiencies of 18 to 19% were developed, with superior radiation resistance (EOL efficiency, 14%) and less fall-off in output when operated at elevated temperatures. The improved performance resulted in significant reduction of area and weight of the solar array, making GaAs/Ge cells cost-effective for the PV system, even though they may cost 5–10 times as much as Si cells. These system advantages resulted in rapid acceptance, especially for commercial communications satellites and defense satellites.

2.2.2 Multijunction Solar Cells

Tandem cells made by stacking cells with appropriate band gaps have higher efficiency. After groups at the National Renewable Energy Laboratory successfully grew gallium-indium-phosphide (GaInP₂) cells over GaAs cells, on the same GaAs substrate, these GaInP₂/GaAs cells were adapted by the space cell community using Ge substrates and production-level OMVPE growth. These dual-junction (DJ) cells eventually attained BOL efficiencies as high as 25.5% and EOL efficiencies of approximately 20%. Additional research led to formation of a PN junction in the Ge substrate during growth of the DJ cell. The resulting triple-junction (TJ) cells have median efficiencies of approximately 27.5% and have been modified to have superior radiation resistance, with EOL efficiencies of approximately

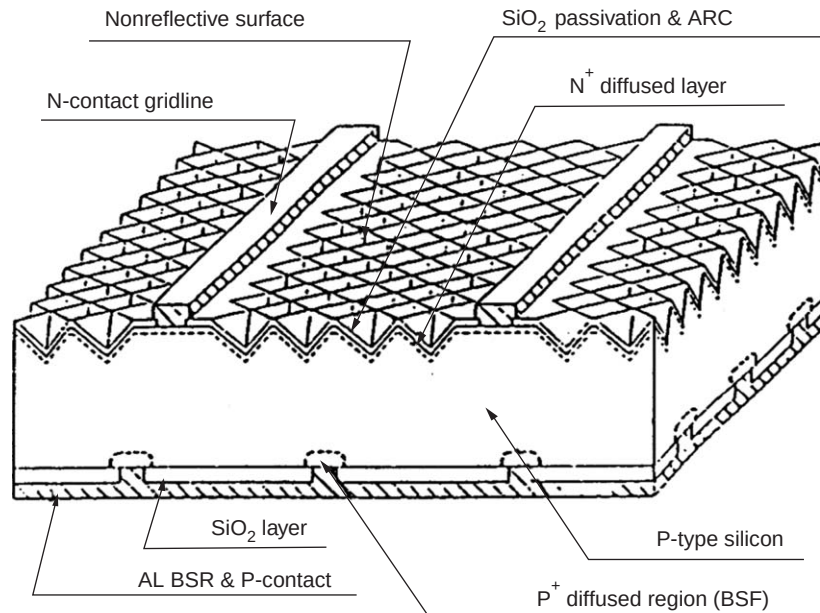


FIGURE 2 Structure of advanced space silicon cell.

23%. Figure 3 shows the cross section of a TJ cell. In addition to the three PN junctions, there were many other layers for passivating interfaces and surfaces and for effectively conducting current through the cell. The growth and properties of all these layers were systematically optimized and controlled by computerized growth programs. Consistency of performance was checked by advanced characterization procedures after each reactor run. The costs resulting from the higher complexity were slightly higher than for GaAs/Ge cells but were offset by the superior performance.

Testing showed that when some multijunction cells in a series string were shadowed, they often degraded. To prevent this shading loss, each MJ cell required its own bypass diode. At first, discrete diodes were used, but now each TJ cell includes a bypass diode formed during OMVPE growth of the cell.

Development of four-junction cells promises efficiencies of 33–35%, but to date efficient four-junction cells have had severe material problems.

2.2.3 Indium Phosphide-Based Cells

Cells containing indium and phosphorus (InP and GaInP₂) have high radiation resistance and show

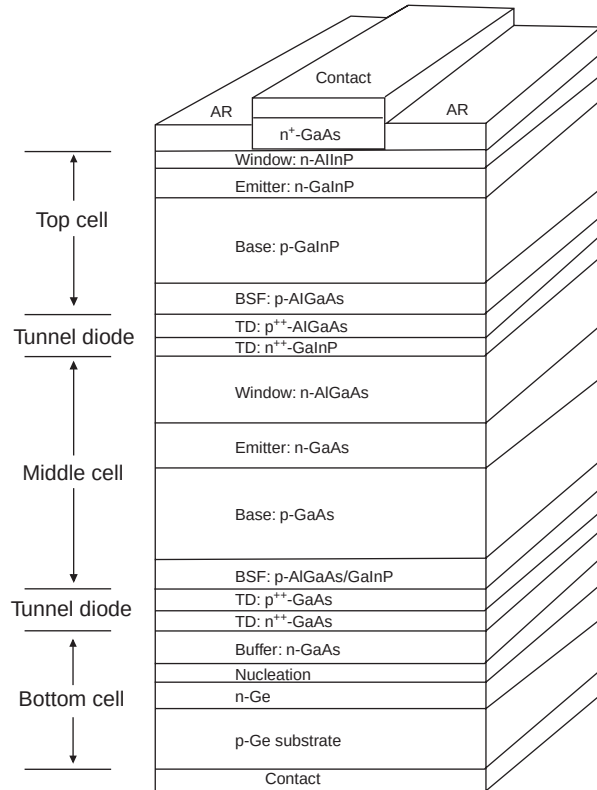


FIGURE 3 Cross section of triple-junction solar cell.

recovery of output after radiation. These cells, especially InP grown on Si substrates, are candidates for orbits with very high radiation levels, including MEO or near to Jupiter.

2.3 Thin Film Cells

Thin film (TF) cells, such as amorphous Si or copper-indium-diselenide (CuInSe₂), have advantages of low mass, easy deployment, and low cost. However, TF cells with efficiencies of 8–10% AMO have required the use of heavy substrates, which offsets their advantages. TF cells may find use for Mars bases or orbital transfer.

2.4 Solar Cells with Concentrators

With slight redesign, mostly of the front contact gridlines, all solar cells have higher efficiency when incident sunlight is concentrated by lenses or mirrors. Near-Earth operation is limited to low concentrations (five suns) because of the difficulty in removing excess heat from the arrays. However, for far-Sun missions, lightweight concentrators have been used successfully to increase the low intensity and low temperatures leading to higher solar array output.

2.5 Adapting to a Variety of Missions

All solar cells can be modified to perform a wide range of different space missions, but the TJ cells have proved to be most effective. Triple junction cells are versatile, being effective for far-Sun journeys and on the Martian surface. They also perform well in near-Sun operation. Closer to the sun, array temperatures are reduced by tilting arrays away from direct sunlight. At temperatures of approximately 120°C, Ge cells become inoperative and TJ cells become DJ cells, still with high efficiency and sufficient output at the increased intensities.

3. PROGRESS IN PHOTOVOLTAICS

3.1 Power Needs

Photovoltaic power systems have evolved to meet the increasing demands for electric power on a wide range of spacecraft. Cell efficiencies, array designs, and reliability have steadily improved (Table II). Until the 1980s, Si solar cells were used most, and since then GaAs/Ge or multijunction (MJ) cells have been used to meet the specified power needs. Most

TABLE II

Chronological Trends in Space Solar Cell Efficiency (%)^a

Cell type ^b	Year									
	1960	1965	1970	1975	1980	1985	1990	1995	2000	2002
A	12	12.2	12.5	12.8	13.0	13.0	13.0	13.0	13.0	13.0
B	13	13.2	13.5	13.8	14.0	14.0	14.0	14.0	14.0	14.0
C	13.5	13.8	14.0	14.2	14.5	14.8	15.0	15.0	15.0	15.0
D						16.5	17.0	17.8	18.0	18.0
E						17.0	18.0	19.2	20.0	20.0
F							20.5	21.8	23.5	25.5
G								23.8	25.5	27.5

^a Air mass zero, 30°C.^b A, silicon, 10 ohm-cm, back surface reflector (BSR); B, silicon, 2 ohm-cm, BSR; C, silicon, 10 ohm-cm, BSR, back surface field (BSF); D, silicon, 10 ohm-cm, advanced design; E, gallium arsenide (on GaAs or Ge substrate); F, dual junction (GaInP2/GaAs); G, triple junction (GaInP2/GaAs/Ge).

satellites have performed successfully with PV power levels of a few kilowatts or less. To handle increased data requirements, communications satellites must deliver power levels up to 15 kW, using MJ cells and sometimes low-concentration mirrors. Even higher power levels are needed for constellations of communications satellites. An example is Iridium, comprising 66 satellites, each delivering approximately 4 kW for a total of approximately 250 kW.

Manned space stations require high power. The International Space Station uses large-area, coplanar back contact Si cells specially designed for deployment of large-area, lightweight arrays. Two wings, each approximately 100 ft long and 40 ft wide, deliver approximately 150 kW, and plans call for adding further wings to produce more than 400 kW. This is approximately 10,000 times the average power available on satellites in the early 1960s. The estimated working life of these cells is 10–20 years.

Some well-publicized spacecraft provide good examples of adapting PV technology to suit the mission. The Mars Pathfinder lander derived approximately 200 W from the unfolded panels and 16 W from the Sojourner robotic rover. GaAs/Ge cells were used, and the output of one cell was monitored to measure the gradual loss of power as Martian atmospheric dust settled on the cell.

Other satellites have approached or landed on comets or asteroids. Deep Space I traveled approximately 100 million miles into the asteroid belt. Its PV arrays consisted of MJ cells with lightweight Fresnel lens concentrators, giving approximately 30-times concentration. In addition to operating on-board equipment, the PV arrays powered an ion engine that

propelled the satellite on its journey. This solar electric propulsion technology is promising for travel within the inner Solar System.

3.2 Interaction with Mainstream Photovoltaics

There has always been close cooperation between groups developing space cells and those developing terrestrial cells. For many years, Si space cells had the highest efficiency, but terrestrial Si cells were developed with significantly increased output. Space PV has been a highly visible showcase for high-efficiency cells. The process procedures for Si space cells were the most consistent available for objective evaluation of the terrestrial solar cell potential of many forms of solar-grade Si. High-efficiency MJ cells designed and operated under incident sunlight concentrated 500–1000 times have promise for large-scale terrestrial PV applications. III–V semiconductor optoelectronic technology has benefited from PV experience with large-scale OMVPE reactors.

4. FUTURE TRENDS IN SPACE PHOTOVOLTAICS

4.1 To 2012

In the next decade, there will be continued and increased demand for PV power on a wide range of missions. Higher power levels will be required, especially for communications satellites and manned stations. Plans call for exploration of Mars or the

moon to prepare for later manned bases. Development of very small satellites used for smart sensing will begin.

Advanced array designs will be developed, and solar electric propulsion will be used for travel to distant destinations. Some satellites will have to operate in high-radiation fields, including ME orbits. For these missions, solar cells with very high radiation resistance will be needed, probably InP or GaInP₂ cells, which have high tolerance to radiation damage as well as the ability to recover later while illuminated by the sun.

Thermophotovoltaic cells developed from III-V solar cell technology may be used to convert heat generated by radioactive isotopes to provide electric power for satellites journeying to distant planets such as Pluto or beyond.

Space-qualified four-junction cells with efficiencies in the mid-30% range will continue to be developed, with effort directed to resolve the severe materials problems involved. The need for higher efficiencies may be relaxed somewhat if transponders can be built to use less power or if electronic circuit designs can further compress the data handled on the satellite.

Current solar cells and arrays and their manufacturing base can probably meet all space power needs in the next decade.

4.1.1 Unmanned Aerial Vehicles

Several applications currently provided by satellites are being planned for unmanned aerial vehicles (UAVs), which include high-flying unmanned airplanes, airships (blimps), or balloons. PV arrays provide propulsion power for the airplanes and airships and power the electronic equipment on the UAVs. UAVs operate at altitudes as high as 100,000 ft for periods of months, and they can be used for surveillance, crop monitoring, or as communications links. The solar arrays on UAVs must be highly efficient and lightweight, and current space-use solar cells are best suited to this application. Current UAV plans call for 50 kW per vehicle, in the next decade increasing to several hundred kilowatts of PV power for constellations of UAVs.

4.2 After 2012

In this period, other power generators, possibly nuclear, will be developed and become available for use in space. It is difficult to predict the needs for space PV beyond 2012, but PV will continue to play a major role. There will be continued extension of most of the previous missions. Manned bases on

Mars or the moon will require large amounts of PV power, and it is an advantage that PV systems are modular, allowing large power stations to be assembled during a series of flights.

The problems in manufacturing four-junction cells will be resolved, providing cells with efficiencies of approximately 35%. Long term, there are exciting possibilities to increase efficiencies perhaps as high as 60–90%. These cells, called “third generation” by Martin Green, will require advanced control over semiconductor material properties and possibly application of nanotechnology. Resolution of these material problems and validation for space use are not anticipated until well after 2012.

After 2012, significant advances will be made in the use of small satellites (called microsats or nanosats). These nanosats will act as smart sensors, operating in autonomous clusters or “swarms.”

In the future, there may be some very large-scale applications of PV power in space. These projects will require significant advances in solar cell and array technology and use of very sophisticated manufacturing methods. Some of these applications may involve assembly in space, and in the distant future solar cells may be manufactured using lunar materials. The following sections discuss some of these large-scale applications.

4.2.1 Space Power Satellites

Peter Glaser proposed placing large satellites in geosynchronous orbits to provide terrestrial power. The satellites would carry very large-area solar arrays, generating many megawatts (or even gigawatts) of electric power. This power would be converted to microwave power, to be beamed to Earth, and converted back to electric power. In principle, the satellites receive full AMO sunlight for almost 24 h per day. However, this advantage is offset by the conversion losses incurred in beaming the power to Earth. Power satellites are more attractive for regions that have low sunlight levels or restricted areas for deploying PV arrays. There are many problems to be resolved, including the cost and complexity of building and maintaining large arrays in space and in controlling and beaming large amounts of power. Also, at this future time, improved terrestrial PV power generators will be widely available at reduced costs, requiring assessment of whether the many problems with space power satellites justify their use to meet terrestrial needs.

Even more ambitious long-term proposals call for installing several bases containing PV solar arrays on the moon. The electricity would be carried by buried

lines to microwave generators beaming power to the earth.

4.2.2 Space Colonies

Colonies living in satellites located at stable positions in space, between the earth and the moon, will require large amounts of power to sustain life and operate equipment. Preliminary plans call for PV power combined with solar-thermal converters.

4.2.3 Space Cities

Arthur Clarke's concept of cities located at GEO altitudes uses space elevators to transport materials and electric power to orbit. This is a very speculative project, requiring materials that can support structures 23,000 miles long. If successful, it would require removing most of the LEO and GEO satellites operating at that time.

5. CONCLUSION

For almost all of the space age, PV has been a prominent and essential key to successful operation of thousands of satellites performing a wide variety of activities. This continuing niche market has shown great resilience, sustained by the manufacturing of space solar cells with significantly increased efficiency and high reliability. Steady improvement in space cells has served as a focus for general research on high-efficiency cells. The programs to develop III-V-based cells demonstrated that large-area, good-quality crystalline layers could be grown in large-capacity OMVPE reactors. Pioneering work led to the use of Ge substrates for these cells. Some of the experience gained with space cells designed to operate at high concentrations has led to possible terrestrial applications. The unique need to deal with the degrading effects of charged particle radiation on space cells has added considerable understanding of

the effect of radiation on semiconductor properties. Current manufacturing technology for space cells and arrays can deliver power for most satellites planned for the next decade and probably beyond. Based on the good record to date, PV power will continue to make major contributions to many exciting space projects.

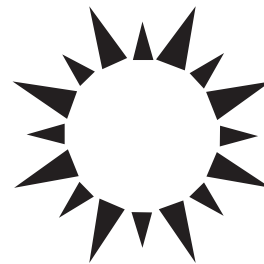
SEE ALSO THE FOLLOWING ARTICLES

Lunar-Solar Power System • Photovoltaic Energy: Stand-Alone and Grid-Connected Systems • Photovoltaic Materials, Physics of • Photovoltaics, Environmental Impact of

Further Reading

- Bailey, S. G., and Flood, D. J. (1998). Space photovoltaics. *Prog. Photovoltaics Res. Appl.* 6(1), 1-14.
- Flood, D. J., and Brandhorst, H. (1987). Space solar cells. *Curr. Topics Photovoltaics* 2, 143-202.
- Green, M. A. (2001). Third generation photovoltaics: Ultra-high conversion efficiency at low cost? *Prog. Photovoltaics Res. Appl.* 9, 123-126.
- Iles, P. A. (1998). From Vanguard to Pathfinder: Forty years of solar cells in space. In "Conference Record of the Second World Conference and Exhibition on Photovoltaic Solar Energy Conversion, Vienna, Austria." Vol. 1, pp. LXVII-LXX. European Commission, Ispra (VA), Italy.
- Iles, P. A. (2000). Future of photovoltaics for space applications. *Prog. Photovoltaics Res. Appl.* 8(1), 39-51.
- Iles, P. A. (2001). Evolution of space solar cells. *Solar Energy Mater. Solar Cells* 68, 1-13.
- Perlin, J. (1999). "From Space to Earth." Aztec, Ann Arbor, MI.
- Rauschenbach, H. S. (1980). "Solar Cell Array Design Handbook." Van Nostrand Reinhold, New York.
- Smith, K. D., Gummel, H. K., Bode, J. D., Cuttriss, D. B., Nielsen, R. J., and Rosenzweig, W. (1963). The solar cells and their mounting. *Bell System Tech. J.* XL11, 1765-1816.
- Smits, F. M. (1976). History of silicon solar cells. *IEEE Trans. Electron Devices* ED-23, 640-643.

Photovoltaic Energy: Stand-Alone and Grid-Connected Systems



TIM MEYER

Fraunhofer Institute for Solar Energy Systems ISE
Freiburg, Germany

-
1. Applications of Photovoltaic Systems and Markets
 2. System Concepts and Special Aspects
 3. Components of PV Systems

Glossary

autonomous electricity supply system Any electricity supply system that operates independently from the grid, be it in remote locations, mobile, or grid-independent applications for which a grid is nearby.

backup generator Controllable source of electricity in photovoltaic hybrid systems (e.g., based on a piston-type engine).

charge controller Electronic device to control the charging and discharging of batteries and thus protect the batteries.

grid-connected photovoltaic system A photovoltaic system that feeds power into the electricity grid using an inverter.

hybrid system Autonomous electricity supply system based on at least two sources of electricity, such as photovoltaic, wind, or diesel generator, always including a battery storage.

inverter Power electronic device to convert direct current (DC) to alternating current (AC).

maximum power point (MPP) The point of operation at which a photovoltaic module provides the maximum power output.

minigrid An AC electricity grid that forms an island and is not or not always part of the regular grid. Minigrids can be used to power small regions or villages or to maintain the power supply in a region when the main grid fails.

solar home system (SHS) Autonomous DC power supply system for light, radio, or TV used for rural electrification of single houses.

stand-alone PV system Autonomous electricity supply system based on photovoltaics as the power source.

Power supplies based on photovoltaics (PV) were first used commercially for space applications and

small appliances, such as pocket calculators and watches. Because of lower prices of PV modules and growing experience in PV system design, higher power ranges have become technically and economically viable.

1. APPLICATIONS OF PHOTOVOLTAIC SYSTEMS AND MARKETS

There are a vast number of applications for photovoltaic (PV) technology, covering all kinds of autonomous systems as well as grid connection (Fig. 1). The power range is extremely wide due to the modularity of PV power, from milliwatts to megawatts. To improve the economics or to use the variety of available energy sources, in a power range of kilowatts and higher so-called hybrid systems are often used. They combine two or more power generators, such as PV, wind turbines, diesel generators, or microhydro generators.

Driven by large market introduction programs, an increasing share of PV capacity has been installed in grid-connected applications. In 2000, these exceeded 50% of the market of the 20 International Energy Agency member countries for the first time (Fig. 2). Due to strong interest in PV technology, further growth of grid-connected systems is expected. However, in the long term the larger economic potential lies in off-grid power supplies. It is estimated that approximately 2 billion people live without electricity from the grid. Autonomous power supplies based on PV or PV hybrid systems are cheaper than diesel systems or grid connection in many of these cases. Socioeconomic and sociotechnical issues are the main barriers to large-scale market takeoff.

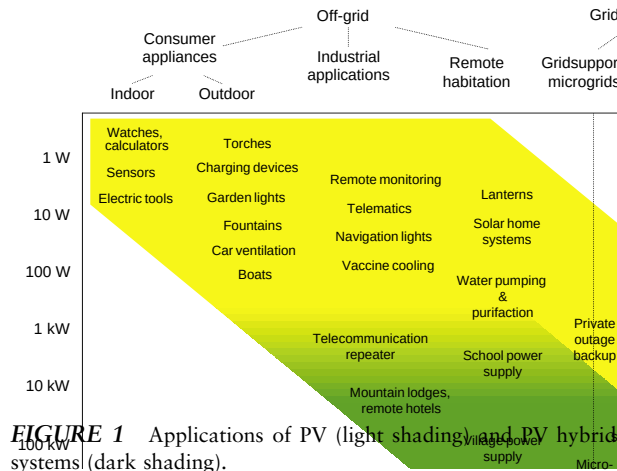


FIGURE 1 Applications of PV (light shading) and PV hybrid systems (dark shading).

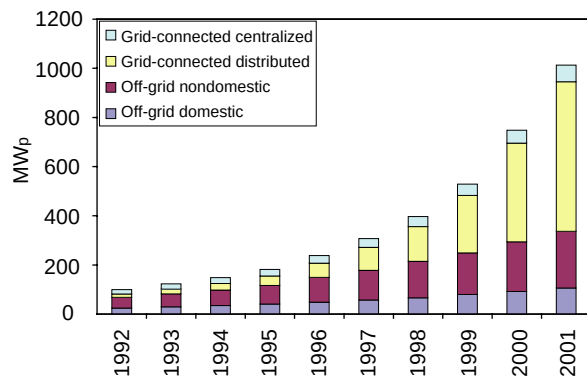


FIGURE 2 PV market development in the IEA countries (Australia, Austria, Canada, Denmark, Finland, France, Germany, Israel, Italy, Japan, Republic of Korea, The Netherlands, Norway, Portugal, Sweden, Switzerland, the United Kingdom, and the United States). Markets for rural electrification are not included. Source: IEA PVPS Programme.

2. SYSTEM CONCEPTS AND SPECIAL ASPECTS

2.1 Autonomous Systems

2.1.1 PV Products and Applications Based on PV Only

Simple autonomous power supplies consist of a PV generator, a charge controller, and a storage unit. However, since batteries have a complex behavior with respect to aging, state of charge, temperature dependencies, etc., the charge controller can have a simple voltage-dependent charge/discharge control or complex algorithms.

Figure 3 shows a solar home system (SHS) as an example of an autonomous power supply system

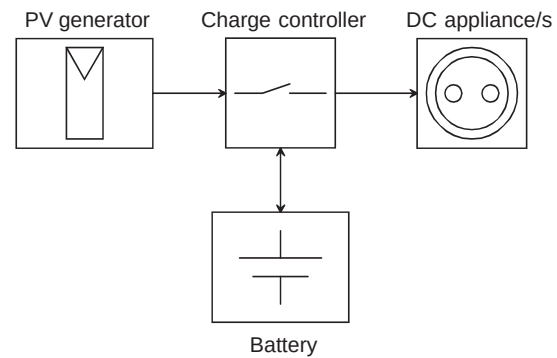


FIGURE 3 Typical configuration of a solar home system.

based on PV only. Typical component sizes are 50 W for the PV module and a 12-V, 70- to 120-Ah lead acid battery; direct current (DC) appliances for this system are compact fluorescent lamps, a radio, and a black-and-white TV set. Typical electricity consumption of such systems is up to 250 Wh per day. In some electrification programs, alternating current (AC) appliances are used in SHSs instead of DC appliances. This requires an additional inverter with a peak power in the range of 150–500 W.

2.1.2 DC-Coupled PV Hybrid Systems

If larger power or energy demand is to be satisfied, additional generators must be introduced into the system. This particularly applies if solar irradiation varies widely throughout the year. To guarantee a reliable power supply during wintertime, large PV generators and storage units are required in these cases. Wind, hydro, or diesel generators are also typically used (Figs 4–6).

Introducing additional components can lower overall electricity cost but also increase the need for energy management and maintenance. Energy management considers the profile of electricity generation from PV or other renewable energy sources, the battery characteristics, and the energy demand of the appliances. To economically optimize energy and battery management, predicting the fluctuating sources (PV, wind, and hydro) can be helpful. The appliances can be either DC or AC. In the latter case, an inverter is used to provide AC electricity. Additionally, an AC bypass from the diesel generator can be used as backup for AC appliances.

2.1.3 AC-Coupled PV Hybrid Systems

Instead of coupling generators and battery storage via a DC bus, hybrid systems can also be AC coupled. The extent to which DC busses are avoided can vary. Fig. 7 shows a system layout in which the

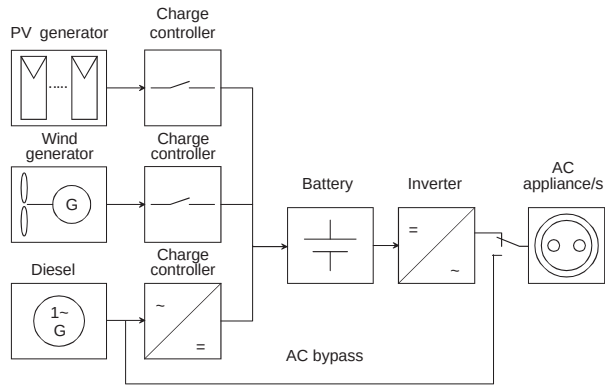


FIGURE 4 Typical layout of a DC-coupled hybrid system.



FIGURE 5 The Rotwand house in the German Alps. Electricity is provided with a PV, wind, and diesel hybrid system in operation since 1992.

PV generator and battery are connected via a DC bus and additional generators are AC coupled to the system. Using the concept of modular system technology, DC coupling is completely avoided and the battery is equipped with extra power electronics (Fig. 8). Modularity adds to the cost of additional equipment and power conversion losses.

In AC-coupled hybrid systems, an island grid is built. There are different ways to achieve this, particularly concerning stability and fast voltage control. In the most common design, one component



FIGURE 6 A telecommunication repeater station powered by a PV hybrid system.

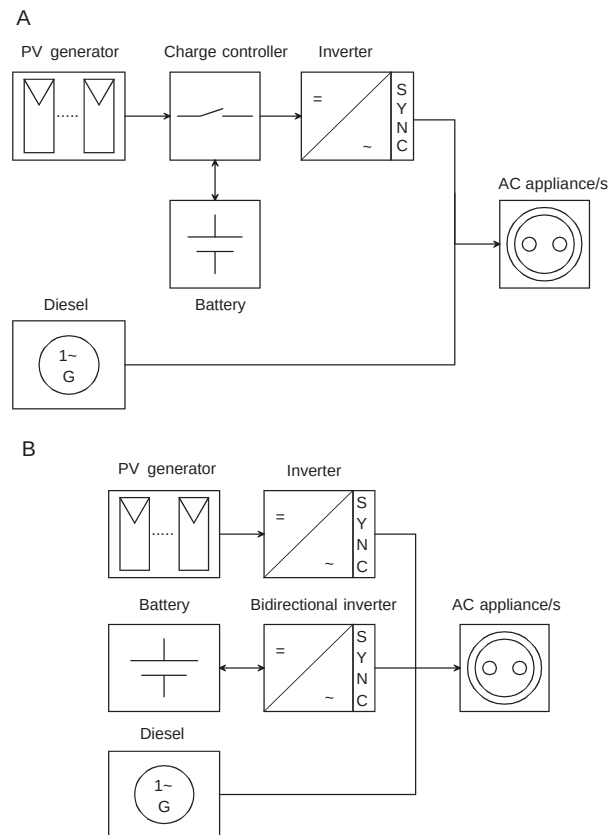


FIGURE 7 Typical layouts of AC-coupled systems (A) with PV generator and battery coupled via a DC bus and (B) following the modular system technology concept.

acts as the voltage source and constitutes the grid, setting voltage and frequency (e.g., 230 V and 50 Hz). The other AC generators then feed in their power, acting as current sources. Reactive power demand (either capacitive or inductive) is provided by the generator acting as a voltage source. The advantage

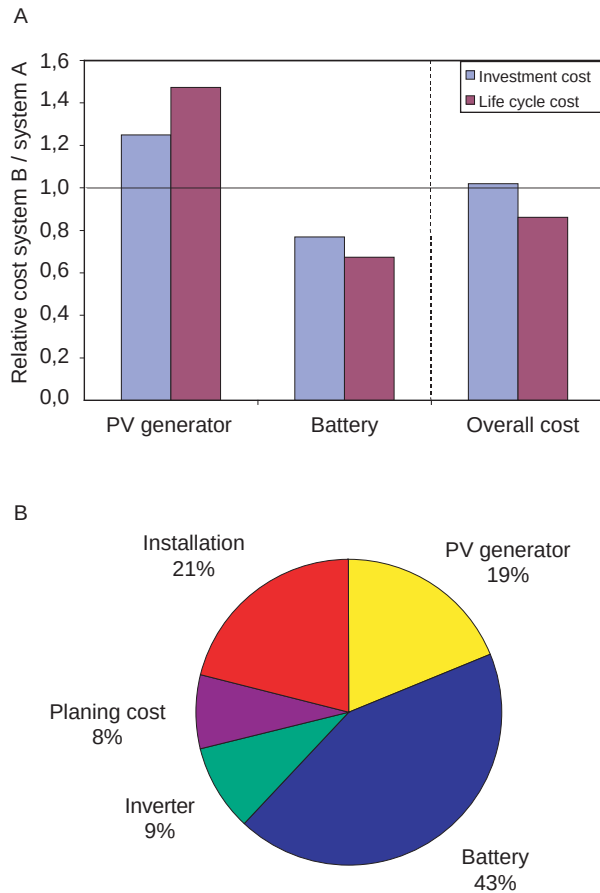


FIGURE 8 (A) Comparison of calculated costs for two systems in Mexico. The systems will provide 1500kWh per year for 20 years. The interest rate is 6%. System A is optimized toward minimum investment cost and system B toward minimum life cycle cost. In system B, costs are shifted toward the PV generator to reduce the battery stress. Although both systems have approximately the same investment cost, system B has 14% less life cycle cost. The more complex hybrid systems are, the more benefit such economic optimization can yield. (B) Shares of life cycle cost for system A. The battery accounts for more than 40% of overall life cycle cost.

is a relatively easy control structure. However, overall system efficiency may be negatively affected, for example, if a diesel generator is used in building the grid and is mainly running idle just to maintain the grid voltage. Furthermore, if fast load or generation changes occur, the constituting component (diesel or inverter) has to cope with them solely and provide all extra power.

As for large electricity networks, new control concepts distribute this burden to several or all generators, which also build the grid (i.e., constitute the AC voltage). Using droop characteristics for frequency and voltage control is one approach to synchronize the voltage control without the need for

very fast and excessive communication between components. The first systems of this type have recently been introduced into the market.

Independent of how fast voltage control is achieved, in DC hybrid systems and AC-coupled systems an overall energy management system determines which components operate and at which power level, thus optimizing energy and battery management.

2.1.4 Main Design Criteria for Hybrid Systems and Advantages and Disadvantages of Different System Concepts

As the number of different types of generators bound together increases, so does the complexity of the system. This also applies if several generators of one type are introduced into the system (e.g., in the case of expansion). Thus, depending on the type of application and the prospect for changing needs in power supply following installation, a number of criteria must be considered during the concept design phase. Some of these are contradicting, so trade-offs must be made.

One of the most expensive components of such systems, the PV generator, has a long lifetime of approximately 20 years. Thus, in order to fully make use of this investment, many systems are designed for such long lifetimes, with the following specifications:

- **Lowest possible life cycle cost:** The cost structure of autonomous power supply systems is often complex and strongly dependent on operation schemes and depreciation times (e.g., 20 years).
- **Stability against load and generation variations:** Daily, monthly, and yearly profiles of the load as well as generation by renewable generation vary widely over time. Furthermore, over the years, the amount of energy needed changes in most applications. For example, additional loads might be connected over the years. In applications such as lighting systems in bus stations, rare blackouts are acceptable. In many others, they are not. The system design must be stable against such variations and have a minimum of costly overcapacity (e.g., during summer).
- **Modularity and flexibility:** If load or generation change too much, the system hardware needs to be changed. Usually, this implies an expansion due to increased energy demand. It must be possible to adapt the system with minimal engineering effort regarding the hardware and the energy management algorithm.
- **Maintenance and repair:** Over the course of many years of operation, maintenance is crucial. The

system must be as easy to maintain as possible. Also, the availability of spare parts may be crucial.

- **Voltage stability and quality of power supply:** Quality of power supply does not only include the availability of power and energy but also the quality and stability of the voltage. Especially in AC-coupled systems, this implies strong demand for fast current and voltage control.

- **Socioeconomics:** In most applications, users interact with autonomous power supplies. Be it in remote houses, village power systems, or for SHSs, without taking sociotechnical and socioeconomic factors into account, the system will fail over the long or short term. This is of particular importance for rural electrification.

- In certain applications such as telematics and signaling, in which blackouts can lead to accidents, the reliability and safety of power supply are of utmost importance.

Many of these demands are contradictory, making optimization complex and susceptible to changed boundary conditions. Thus, new system designs have been developed to more easily and flexibly optimize systems during the planning phase as well as during operation. The designs for energy management constitutes the most important system characteristics and are thus described in more detail.

2.1.5 Energy Management in Off-Grid Systems

Energy management is usually achieved by one central intelligence. All system parameters, such as generator and storage sizes, storage characteristics and charging/discharging strategies, weather forecast (e.g., for PV and wind), and user profiles are fed into the system before operation. Using these parameters and actual measured values of voltages, currents, temperature, PV generation, etc., the system operation is optimized. In addition to standard software, fuzzy logic and artificial neural networks are also used. However, expansion or changes to the system require changes to these system parameters. Furthermore, the system operator requires data that are often not well specified on data sheets, particularly for the battery and the backup generators. Thus, determining a proper operating strategy is difficult.

Alternative methods from automation and control in industrial applications or from management of distributed generation try to decentralize as much intelligence as possible, combining a central control unit with one decentralized control unit per generator, storage, or load. Corresponding software concepts can vary from hierarchically controlled

communication and interaction to software agents. In all these cases, component information is stored in a device preferably provided by the component manufacturer. Using measured values, this device is also able to compute cost functions and the availability of the generator or storage (e.g., using the filling state of the diesel tank and maintenance intervals). For these calculations, some information may be provided by the central control unit, such as weather forecasts or the cost for maintenance personnel to visit the site of operation. All cost functions and limiting conditions from the components are then used to optimize the system operation (e.g., using virtual trading schemes). Thus, there is no need to change the energy management parameters when changing the system design. New components are registered and removed components are deregistered automatically and provide data and cost functions in a generic format. The virtual stock market takes into account all registered players at a certain time. Such flexible systems are currently under development and not available on the market.

2.1.6 Cost Structure of Autonomous Power Supply Systems

Autonomous power supply systems often have long lifetimes or depreciation times of up to 20 years. Thus, to develop a mature system design, life cycle cost needs to be considered rather than investment cost. The following are especially important to consider: the complex dependency of aging of the battery on the operating conditions (batteries need to be replaced every 4–8 years), repair or replacement of electronic components, costly maintenance in remote locations, and the cost of financing and penalties for electricity demand that may not be met (if allowable). Because weather profiles differ from year to year, the technical optimization should be based on long time series of 20 years. The difference between systems optimized toward investment cost and those optimized toward life cycle cost is shown in Fig. 8. The main difference between system designs is that, if optimized toward minimum life cycle cost, the battery costs are reduced by appropriate design and operation strategies and by the expanded lifetime or more efficient use of the battery. It is important to note that in autonomous systems the PV generator is not the largest cost factor and that the overall life cycle cost of the system is typically approximately three times more than the pure investment cost (considering interest payments, replacement of components, maintenance and repair, fuels, etc.). When designing an autonomous power

supply, the choice of power generators and their sizes should be determined only with respect to overall lifetime costs. All options should be considered. Nevertheless, in many cases PV or PV hybrid systems are the most cost-effective solution.

2.1.7 Socioeconomic Aspects of Autonomous Systems and Rural Electrification

Autonomous power supplies are the most promising technology for the 2 billion people living without electricity today. During the past few decades, in all regions of the world many projects have demonstrated the technology and its viability. However, a large number of these projects were not sustainable due to equipment failure after a few years, payments not being made by the rural customers, or electrification programs not being accepted by the population as a viable alternative to grid connection. From such projects, many lessons have been learned providing valuable feedback for the design of components and systems as well as business models, infrastructure programs, and policy.

In rural electrification, it is of utmost importance to include cultural aspects in technology and business model design. Sustainable projects and thus sustainable business require not only a screening phase of the respective region and a planning and implementation phase, but also sociotechnical monitoring during operation and ongoing analyses to adapt the strategy. Companies and staff have to learn the specifics of rural electrification and the sociocultural backgrounds of the people, and they must understand that rural electrification is a process. The following issues are crucial:

- There is often a gap between the energy demand and the supply by the PV or PV hybrid system after a short period of time. Once electricity is available, people buy new appliances and consume more energy than expected. This particularly applies if there is little “ownership” for the system (e.g., if a village power supply is used by many users). In this case, the cause (too high energy consumption) and the effect (blackout) are not easily linked to the individual user. In any case, if demand is too high and cannot be met, dissatisfaction is created, possibly leading to stoppage of payments, etc. Although difficult, the future electricity consumption needs to be evaluated before the system is installed.
- Due to erroneous handling of components and systems, these systems degrade or break more quickly.
- Replacement of components is time-consuming and expensive. It is vital to develop a minimum

density of systems within a certain area and to install an adequate infrastructure. This includes measures for capacity building, such as training of technical staff and users.

- Lack of financing, especially after several years of service, is often experienced. There might be clear reasons, such as dissatisfaction of customers, but even if the systems run properly, payments tend to decline significantly after several years. Financing schemes and business models need to take this into account.

2.2 Grid-Connected Systems

Grid-connected systems constitute the largest market share for photovoltaics. Major driving forces are market introduction programs in Japan, Germany, and the United States. These systems are targeted at households. Typical system sizes range from 1 to 5 kW. However, there are commercial actors and many systems more than 10 kW and up to several megawatts have been built (e.g., utility owned or joint ownership). The power of inverters that is required to feed the PV electricity into the grid ranges from 100 W to more than 100 kW.

The smallest inverters are module integrated ones, with each single PV module having its own inverter. Main advantages are lower losses at partial shading and lower mismatch losses between modules due to individual maximum power point (MPP) tracking and less demanding cabling because standard AC cables can be used. DC cabling and switches are more expensive because in the case of faults arcs can cause fire. On the other hand, maintenance and repair of many small inverters are more difficult and communication between inverters (e.g., to determine the overall output) requires higher effort. Cost benefits have not yet been realized. It has also been reported that large numbers of module integrated inverters influence each other through increased harmonics on the AC line. Today, few systems are built with module integrated inverters.

The most common system is shown in Fig. 9. A number of PV modules are connected in series to form a string, typically with output voltages at the MPP of 700 V. One or more parallel strings are connected to one inverter to feed the power into the grid. Depending on regulations and inverter power, either one-phase or three-phase inverters are used. Often, large systems are built using many inverters in the range of a few kilowatts. Again, the modularity of PV electricity is apparent.

In addition to the modular concept shown in Fig. 9, for large PV systems, large inverters are used with up to 250 kW or even 1 MW per unit. Both designs, few large inverters or many medium-size inverters, are motivated by cost and neither is superior by principle. The first makes use of the economies of scale, and the latter makes use of the economies of quantities.

The generation profile of PV systems strongly depends on location. In the Northern Hemisphere, most of the energy of PV systems is yielded at partial load, as shown in Fig. 10. Thus, high inverter

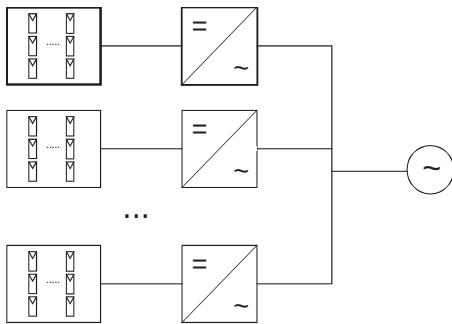


FIGURE 9 Configuration of a grid-connected PV system consisting of several strings.

efficiency at partial load is mandatory. Furthermore, PV inverters with a nominal power rating between 10 and 20% below the peak power of the PV generator are typically used. The inverter is able to handle short periods of overload situations and the energy loss is small compared to the cost reduction due to the smaller inverter.

The performance ratio is used as a measure of the quality of PV systems that is site independent with respect to insolation. The performance ratio is defined as the annual AC energy output of a PV system divided by the annual global insolation in the module plane weighted with the nominal PV module efficiency. Good systems achieve a performance ratio of approximately 0.75. For larger commercial power plants, values of approximately 0.80 should be obtained. The performance ratio aggregates most component and system characteristics: mismatch losses between modules and strings, real module efficiency compared to nominal efficiency, inverter efficiency, and quality of MPP tracking.

2.2.1 Safety

Technical standards, and thus system layouts and inverter concepts, differ from country to country.

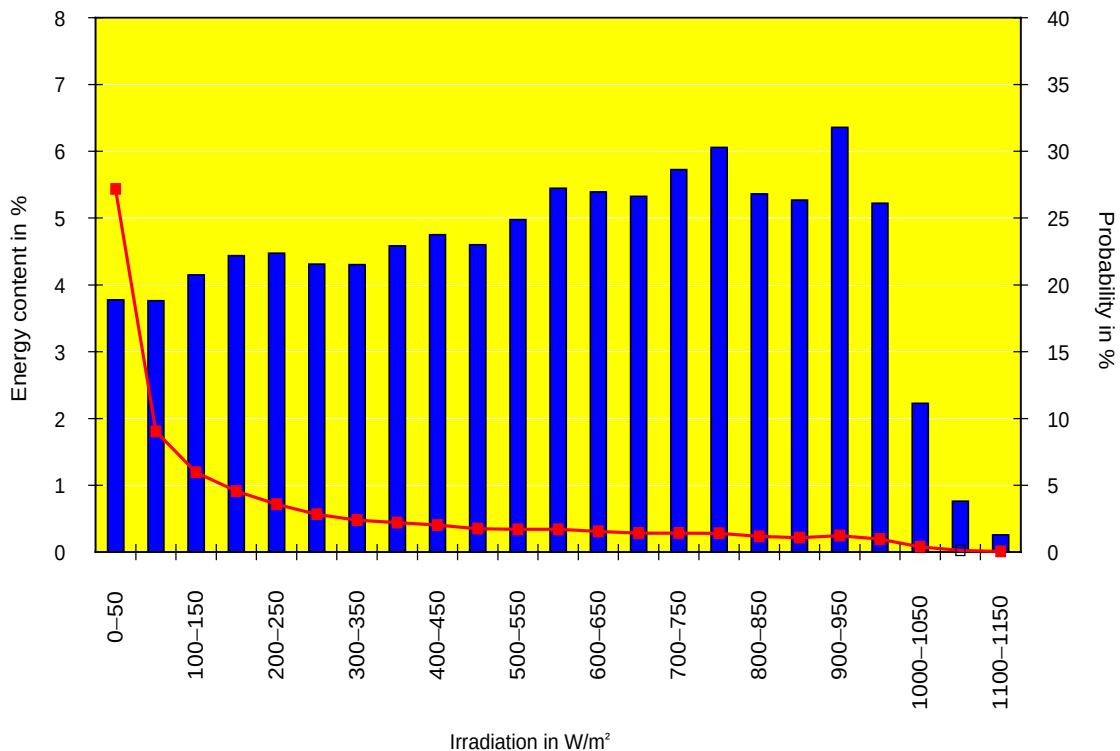


FIGURE 10 Distribution of the energy yield of a PV system in southern Germany. The bars show the energy content for each class of irradiation, and the line shows the probability of the occurrence of this class with respect to time. Most of the electricity is generated at partial load.

Whereas in the United States earthing of the generator is mandatory, in European countries the use of isolated modules (safety class II) is standard. This also allows different inverter designs, particularly transformerless inverters, which have been increasingly used due to their lower weight and higher efficiency.

In addition to earthing, protection against islanding as well as fault current detection in systems with transformerless inverters are necessary, and these are treated rather differently. Islanding is a phenomenon in which during a blackout distributed power sources such as PV systems keep a grid segment—an island—live. This could pose a danger to users, maintenance personnel, and equipment. However, the prerequisite for islanding is that active and reactive power supply by the distributed sources and the active and reactive power demand in the segment be very close. Thus, to prevent islanding it is sufficient to measure voltage levels and frequency and automatically disconnect the inverter if the standard margins are left. If the maintenance personnel follow the safety rules, PV systems do not pose any danger. However, often much more complex methods are used to prevent islanding. With standardization of PV system technology, this might change.

2.2.2 Metering and Cost

To meter the electricity generated, either an extra meter is installed or the electricity is fed, for example, into the house wiring that supplies existing loads (net metering). In the latter case, if PV generation exceeds the actual load, either the meter runs backwards or an extra meter measures the excess electricity.

The generation cost of PV electricity in grid-connected systems varies between approximately 30 €c in sunny regions and 60 €c in northern Europe if good practice is assumed. Unlike in autonomous systems, the PV modules contribute the largest share of life cycle cost. Further cost reduction, however, is achievable using inverter technology and better installation practices.

2.2.3 Building Integration Aspects

Solar energy by nature is a distributed energy source. In order to optimally use existing infrastructure and not unnecessarily seal additional areas, building integration is the best option for grid-connected PV. The PV modules can also provide the following:

- Weather protection (wind and water)
- Sun protection (glare and heating of buildings)
- Use of daylight

- Noise protection
- Act as a blind
- Shielding of electromagnetic fields
- Aesthetics

As such, there are a large number of PV module technologies involving isolating and nonisolating modules, different mounting principles, different module and cell colors, etc.

The cost of integrating PV in rooftops and facades and that of the required standard material can be regarded as a credit. Particularly for facade integration, for which costly material such as marble or metal constructions are often replaced, this can compensate a large share of the overall system costs. This remains true even though facade-integrated PV modules have a significantly lower energy yield due to the nonoptimal tilt angle. In northern Europe, the energy yield is reduced by approximately 30–40% (Fig. 11).

For roof and facade integration, shading of PV modules and operation temperature are of special importance. If one module of a string of modules is shaded, the output of the whole string can be dramatically reduced, depending on the connection scheme and the use of bypass diodes. Thus, particularly for PV facades, which often experience complex shading situations from neighboring buildings or trees, the system must be designed carefully. System parts that are shaded during certain times should have a separate inverter and should be made less susceptible through the use of additional bypass diodes.

Because temperature affects the performance of PV modules, rear ventilation by natural convection is

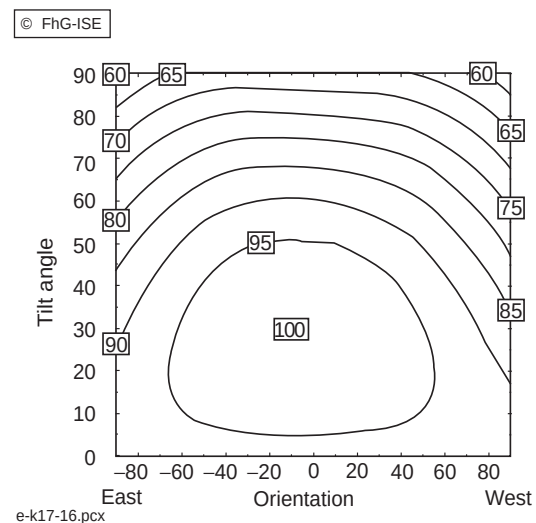


FIGURE 11 Maximum annual PV module yield as a function of orientation and tilt angle in Freiburg, Germany (48°N, 7.8°E).

another issue to consider for building integrated systems. The lower the temperature, the higher the yield. The difference in performance between PV modules with good ventilation and those with bad ventilation can be up to 10%.

2.2.4 Distributed Generation Aspects

Building integration of PV modules can provide significant additional value, as described previously. Another dimension of added value can be tapped if the service provided to the grid is considered. However, improperly integrated PV can also pose additional burden to the grid and thus cause additional cost. Possible factors are as follows:

- Correlation of PV generation and load (reduced peak load in distribution grids)
- Increased security of supply in weak grids (continued operation during blackouts in remote areas)
- Reduced distribution losses in cables and transformers
- Reduced load of transformers, cables, switches, etc. leading to reduced aging
- Deferral of investments for capacity expansion of existing grids
- Active compensation of reactive power or harmonics
- Reduced risk due to short lead times for planning and modularity of PV technology

The actual added or reduced value of grid-connected systems has not been studied. However, several case studies show that at least part of this potential can be tapped.

3. COMPONENTS OF PV SYSTEMS

3.1 PV Modules

Crystalline silicon solar cells comprise approximately 90% of the market share of PV modules. These PV modules typically consist of 36 cells, but other configurations can be bought off the shelf. The current-voltage curve of a 36-cell module fits very well to 12-V lead acid batteries, so simple autonomous power supplies can be built. However, the exact number of 36 cells is more historically motivated. With improved cell technology, even at higher ambient temperatures the module voltage would still be high enough to fully charge a lead acid battery if 32 cells were connected in series. Depending on climatic conditions, even less cells could be used.

The yield of a PV system is usually measured in kilowatt-hours per kilowatt peak (kWh/kW_p) and depends on several factors:

- Cell and module technology used
- Location/climatic conditions
- Tilt angle and orientation
- Mechanical installation (facade or roof integration, freely ventilated modules etc., influencing module temperature)
- Efficiency of system components (MPP tracking and conversion efficiency)
- Type of system

The area needed for 1 kW of PV power mainly depends on the cell and module technology used. Module efficiencies vary from several percent (for some thin-film modules) to approximately 16 or 17% for monocrystalline silicon solar cells. Typical module sizes for grid connected systems range from 50 to more than 300 W. Especially for solar lanterns and lighting, modules in the range of 10–20 W_p are available. Because 36 cells per module are standard, smaller modules are made from cells that are cut into half or quarter cells. Thus, small PV modules with crystalline silicon solar cells are more expensive. This does not apply to thin-film modules.

The performance of PV modules depends on their operation temperature. Depending on the technology used, the temperature coefficient is between -0.2 and $-0.5\%/K$. Thus, the type of installation and degree of rear ventilation are important.

PV modules are rated in watt peak under standard test conditions (STCs). The STCs are defined as 1000 W/m^2 insolation with a AM solar spectrum of 1.5 and 25°C module temperature. However, these conditions rarely occur in real systems. Thus, PV generators can actually yield significantly less or more energy than rated, depending on the location, climatic conditions, etc. However, there is no widely accepted alternative rating model that takes into account more realistic reporting conditions. Furthermore, there is relatively large uncertainty of the measurements at the manufacturing plant and on site. Spectral mismatch, temperature distribution, etc. comprise approximately 3–5% of uncertainty. Overall deviations in systems measurement can be 5–10%.

Whereas in grid-connected systems usually all energy can be fed into the grid, in autonomous systems part of the energy yield in summer cannot be used due to limited storage capacity. Thus, the tilt angle of PV modules in autonomous systems is usually optimized for large energy yield during winter. The overall yearly energy yield is reduced.

Fig. 11 shows the maximum energy yield of PV modules as a function of orientation and tilt angle for the Northern Hemisphere (48°N , 7.8°E).

3.2 Batteries

Batteries are a central component in autonomous power supply systems based on PV or PV hybrids. As discussed previously, batteries contribute the largest share of cost of autonomous systems if life cycle cost is considered. Thus, except for small, portable systems, lead acid batteries are the most common technology used. They are the most cost-effective and have good electrical characteristics (e.g., energy efficiency in PV systems can be as high as 85–90%). Several different technologies for electrodes and electrolytes are available. Thus, for all applications suitable products can be chosen with a wide range of electrical and lifetime characteristics.

For reliable and economic operation of autonomous systems, the charging/discharging principle and system operation strategies are highly relevant in order to make full use of the installed capacity and reduce aging of the battery. In particular, most IV (current phase, voltage phase) charging strategies in autonomous systems rarely fully charge a battery if non-optimal conditions are assumed (e.g., during periods of reduced PV generation). Deep discharge in winter or during longer periods with little generation from PV or wind generators accelerate aging processes.

Depending on the application, different classes of systems can be defined that determine how the battery is stressed (Fig. 12). In addition to the dimensions of the system, the choice of battery type and operation strategy is strongly influenced by these classes.

3.3 Charge Controllers

Charge controllers ensure that the battery in autonomous systems is treated as well as possible, particularly that deep discharge and overcharging are avoided. Furthermore, some devices provide information about the battery state of charge. Modern charge controllers use charging and discharging procedures that take into account the battery temperature and the charging and discharging currents over time. Today, IV charging is state of the art, but new charging strategies are being developed [e.g., IVI (current phase, voltage phase, current phase) charging].

There are two basic designs for charge controllers in PV applications: series or parallel regulators. A series regulator adds a voltage drop between the PV module and battery, whereas a parallel regulator draws current from the PV module bypassing the battery (Fig. 13). Hybrid configurations have also been developed. In all cases, voltage regulation can be performed statically or can be achieved as an average by pulse width modulation. The advantage of the latter is reduced power losses in the

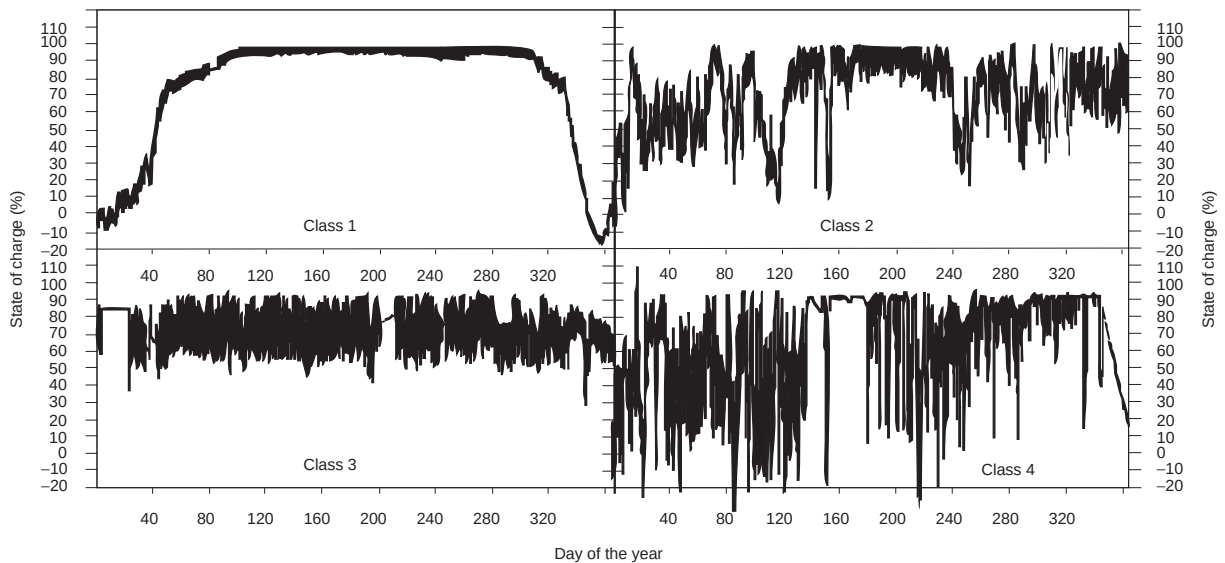


FIGURE 12 State of charge of the battery during a year for four typical operating conditions. The annual energy turnover ranges from approximately 20 times the nominal capacity in class 1 to 180 times in class 4. This demonstrates the different requirements for the batteries under different system designs.

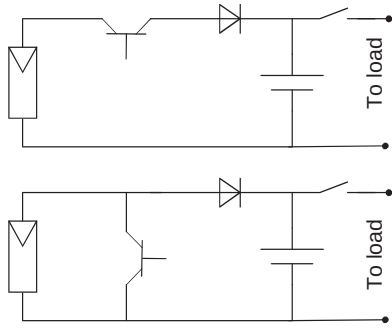


FIGURE 13 Basic principles of charge controllers: series (top) or parallel (bottom) regulator. To avoid deep discharge, the load can be disconnected. A series diode prevents reverse power flow through the PV generator.

semiconductor switch for the microcycles in the battery current and electromagnetic interference (EMI).

The most important functions of charge controllers are to

- Avoid high voltages during charging of the battery (temperature compensates cutoff voltage by -4 to -6 mV/K per battery cell)
- Avoid deep discharging of the battery
- Avoid extended periods without full charging of the battery
- Slightly overcharge once a month to ensure that the battery is fully charged and to promote gassing if applicable

Since some of these aims are contradictory, suitable compromises must be found (e.g., between avoiding high voltages to reduce corrosion and occasionally overcharging the battery to reduce sulfation).

Charge controllers should have a low internal consumption, high efficiency (96–98%), and make reverse poling impossible. They should withstand overvoltages (at least twice the open circuit voltage of the solar generator) and ambient temperatures of 0 – 50°C . Some charge controllers use DC/DC converters to adapt PV module and battery voltage. They can also perform active MPP tracking, whereas with most charge regulators the PV module is operated at the level of the actual battery voltage.

Additional functions, such as starting backup generators and stirring the electrolyte, might be added for systems of higher complexity. In any case, providing information about the state of charge or even state of health of the battery is crucial in order to enable the user to learn how to operate the system effectively, to establish optimal energy management,

and to avoid detrimental deep discharging of the batteries.

3.4 Inverters

Inverters are used to convert the DC output of PV or a battery to AC electricity, either to be fed into the grid or to form an island system. There are many different types of power electronic topologies used in the market. Early PV systems were equipped with thyristor inverters, which were commutated by the grid. Due to their poor voltage and current quality (such inverters have very large harmonics), they have been replaced by self-commutating inverters (with IGBT or MOSFET as semiconductor switches). Only in the 100-kW power range are thyristor inverters still used. The first self-commutating inverters performed sine modulation of the output current on the primary side and had a 50-Hz transformer to adapt the voltage level to grid voltage. Such inverters still comprise a significant market share. Some manufacturers offer devices with high-frequency transformers instead, thus reducing the efficiency and the cost. In countries in which earthing of the PV generator is not mandatory, transformerless inverters are increasingly used. They are lighter and more efficient, but particular care must be taken with regard to EMI and fault current detection.

Whereas the first PV inverters were often adaptations of existing inverters for electrical drives, today's devices are manufactured taking into account the complexity and demands of specific PV applications. Inverter efficiency is generally important in PV applications but of particular significance at partial load because the bulk of the energy is yielded at partial load. Furthermore, with a large-area PV generator coupled to the DC side of the inverter and the public grid on the AC side, stricter standards have to be met with regard to harmonics and EMI.

Design criteria and functions of PV inverters are as follows:

- Efficiency: well above 90% efficiency at 5% of nominal load
- Cost
- Weight and size: inverters with a 50-Hz transformer typically weigh more than 10 kg/kW, sometimes causing a problem with handling
 - Voltage and current quality: harmonics and EMI
 - Overload capability: approximately 20–30% for grid-connected inverters and up to 200% for short-term overload for island inverters

- Precise and robust MPP tracking (reliably finding the overall MPP in partial shading situations)
- Supervision of the grid, safety/ENS (Einrichtung zur Netzüberwachung)
- Data acquisition and monitoring

To optimize inverters following these partially contradictory aims, the design of the power stages and control algorithms are increasingly more integrated. To implement suitable control schemes, digital signal processors are increasingly used.

To account for the fact that PV electricity is yielded following a typical distribution, in Europe a weighted efficiency measure is used to aggregate the efficiency curve of inverters—the European efficiency. Good inverters in the kilowatt power range have a European efficiency of 92–96% and up to 97% for large central inverters.

Inverters also perform MPP tracking in order to optimally operate the PV generator. Many different algorithms are used for MPP tracking for the following:

- Precision: high precision of MPP tracking requires high-precision measurement components
- Finding the global maximum power output in the case of partial shading, when a local maximum can occur
- Quickly adapting the MPP to changes in insolation (e.g., if clouds pass by).

3.4.1 Inverters for Island Systems and Minigrids

Inverters for island systems and minigrids are very similar to inverters for grid connection with regard to the power stages. Most inverter topologies are suitable for island operation as well. The major prerequisite is the capability to provide reactive power. The actual design of the power stages still differs significantly because in island systems the inverter requires a much higher overload capability. Furthermore, fast voltage and current control and higher dynamics of load profiles require appropriately dimensioned power stages.

Thus, there are major differences concerning control algorithms for voltage and current control. Particularly when several inverters are to be connected in parallel in island systems or when inverters are to be combined with other AC generators, achieving voltage stability is a complex task.

3.5 Backup Generators

In PV hybrid systems, backup generators are used to provide additional electricity in times of low insola-

tion, to increase system availability, and to reduce overall electricity cost. In most medium or large power applications, diesel generators are the cheapest backup solution (backup capacity of 5 kW or more). In many countries, subsidized prices for diesel contribute significantly. However, the major disadvantages of piston-type backup generators are their need for maintenance every few hundred hours of operation and bad efficiency during operation at partial load, which is the common operation mode. In remote areas, access to the systems is often difficult or time-consuming, thus imposing extra cost for maintenance and fuel transport. The availability of spare parts and competent maintenance staff can be a major problem in developing countries in which diesel or PV–diesel hybrid systems are used for electrification of villages.

A major problem regarding relatively small hybrid power systems is the lack of suitable backup generators in the power range of 1 kW or less. Thus, oversized piston-type generators are sometimes used for applications in which 1-kW backup capacity would be sufficient. Due to the poor efficiency of diesel and Otto engines under partial load conditions, the overall system efficiency of such systems is not satisfactory.

In small hybrid systems, thermoelectric converters with a backup capacity of 10 to hundreds of watts are also used. In the medium term, fuel cells are an interesting alternative to small diesel or thermoelectric units because they promise high efficiency and low maintenance cost. However, the problem of hydrogen supply needs to be resolved.

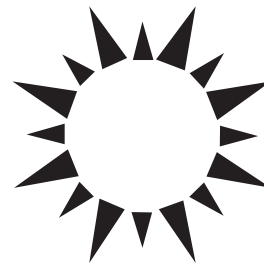
SEE ALSO THE FOLLOWING ARTICLES

Batteries, Overview • Photovoltaic Conversion: Space Applications • Photovoltaic Materials, Physics of • Photovoltaics, Environmental Impact of • Solar Cells • Solar Energy, History of • Solar Thermal Power Generation

Further Reading

- Erge, T., and Sick, F. (1996). "Photovoltaics in Buildings—A Design Handbook for Engineers." James & James, London.
- Gabler, H., Preiser, K., Bopp, G., Sauer, D. U., and Schmidt, H. (1998). Energy storage in photovoltaic stand-alone energy supply systems. *Prog. Photovoltaics* 6, 271–291. [Special issue].
- Luque, A., and Hegedus, S. (eds.). (2003). "Handbook of Photovoltaic Engineering." Wiley, New York.

Photovoltaic Materials, Physics of



BOLKO VON ROEDERN

National Renewable Energy Laboratory
Golden, Colorado, United States

-
1. Solar Cell Operation
 2. Crystalline Silicon
 3. Thin-Film Silicon
 4. Polycrystalline Thin Films
 5. Concentrator Schemes and III–V Solar Cells
 6. Emerging Schemes
 7. Future Potential (Physical Limitations of Materials)

Glossary

amorphous silicon An amorphous semiconductor that has allowed fabrication of moderately high-efficiency thin-film solar cells using a deposited layer of silicon-hydrogen alloy. It lacks crystallinity or the exact periodical arrangement of atoms.

buffer layer A thin layer inserted near the junction of a solar cell to enhance its performance, mostly by enhancing the voltage of a solar cell.

cadmium telluride A thin-film semiconductor that has allowed fabrication of high-efficiency thin-film solar cells.

copper indium diselenide A chalcopyrite semiconductor that has allowed fabrication of very high-efficiency thin-film solar cells.

diode A rectifying semiconductor junction. If designed appropriately, it will deliver electric power upon illumination and function as a solar cell.

silicon The most common material used to build semiconductor devices and solar cells.

solar cell A solid-state semiconductor device used for the direct conversion of light into electric power (i.e., photovoltaic energy conversion).

At the beginning of the 21st century, the direct conversion of sunlight into electricity (solar electricity) via semiconductor devices called solar cells or photovoltaic modules was commercially established. The total number of photovoltaic modules produced worldwide (approximately 500 MW in 2002) was much too small to be considered energy significant.

However, energy significance can be expected between 2010 and 2015, when annual production levels of several 1000 MW per year are reached, based on expected annual growth of 20–30%. Assuming a conversion efficiency of 10%, 1000 MW of photovoltaic module power will cover an area of 10 km² or approximately 4 square miles. In a sunny location, an area of approximately 160 × 160 km (approximately 25,000 km² or 10,000 square miles) would be required to generate from solar power all the electricity currently consumed by the United States. Therefore, a tremendous industrial scale-up must occur to accomplish these goals. The material candidates likely to be used for achieving such commercial potential and the associated physics that relates to solar cell performance and performance limitations are discussed in this article.

1. SOLAR CELL OPERATION

1.1 Construction of the Solar Cell Device

A solar cell is a device that allows the collection of excess electric carriers, called electrons and holes, because of a built-in potential. When semiconductors are used to build such a device, the operation is determined by the bandgap [E_g , usually given in units of electron volts (eV), or voltage multiplied by the charge of the electron e] or energetic separation of the conduction and valence bands. Light is absorbed when the quantum energy of the light (the photon energy) is greater than the bandgap. Under the influence of the built-in potential and, to a lesser degree, gradients in the carrier concentration, carriers will flow from one side of the cell to the other. Electrons will move along the conduction band edge, and holes will move along the valence band edge. Figure 1 represents a solar cell in its most generic form. The schematic depicts the physical composition of layers, using terminology that typically describes

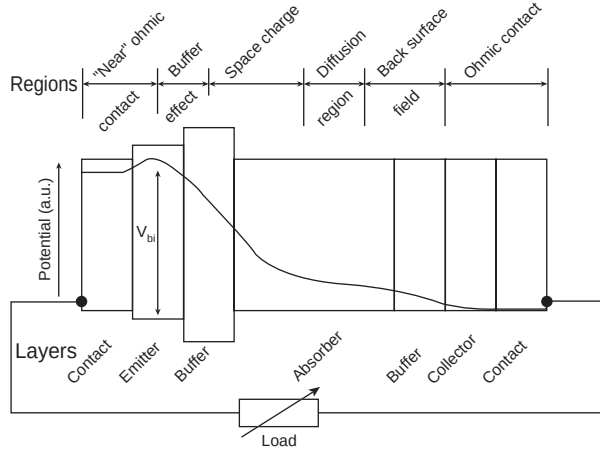


FIGURE 1 Schematic of a p-i-n diode. The potential is shown for no-load (i.e., open-circuit) conditions.

thin-film solar cells. The terms along the top of the schematic indicate the electronic regions of the cell. The built-in potential, V_{bi} , will determine the maximum voltage that the cell could theoretically deliver. For a homojunction cell such as a crystalline silicon cell, the emitter and collector are usually formed by heavily doping the silicon p-type and n-type (as shown for the right-hand collector in the schematic). However, in heterojunction cells, the bandgaps of the materials making up the junction may vary (as shown for the left-hand side, i.e., the emitter in the schematic). For a homojunction cell, the maximum value of V_{bi} is approximately, but less than, E_g/e . In heterojunction solar cells the determination of V_{bi} is less straightforward because the bandgaps of the emitter, buffer, and absorber layers are different. In this instance, in addition to the value of E_g of each layer, the relative alignment of the conduction or valence band edges depends on the “affinity” or work function of each layer. However, in all cells, the potential profile is a smooth continuous curve passing through the device. Emitter and collector layers with low conductivities may not have enough lateral conductivity to collect the currents at a metal gridline. In these instances, the insertion of a transparent conductive layer is required. SnO_2 , ZnO , or indium tin oxide layers are used as transparent conductors. Often, this layer has a double function, acting as an antireflection coating in addition to providing the electrical contact.

The potential profile through the cell will also be determined by the load. Figure 1 depicts the device near open-circuit conditions. When one device contact is shorted to the other one, they have to come to

the same potential, forcing a lowering of the electrical potential on the left-hand side down and/or rising on the right-hand side. Traditionally, much analysis went into identifying collection losses within a specific layer or region of a cell. Regions are often characterized by the material properties of the layer making up most of it. It is important to realize, however, that the distinction of such regions in terms of layers alone is somewhat arbitrary. Manipulation of the physical construction of one region of a solar cell or merely a change in the operating condition of the solar cell device will affect what occurs in adjacent regions.

1.2 Solar Cell Operation

The device shown in Fig. 1 is basically a p-i-n diode, where the emitter and collector doping define the junctions. The cell absorber layer can be lightly doped, undoped (i = intrinsic), or, in some instances, show a naturally prevailing doping that cannot be easily controlled by adding chemical dopants. If a bias is applied to such a diode in the dark (replacing the “load” in Fig. 1 with a power source), the current flow is blocked for one polarity of the applied voltage (reverse direction) and increases exponentially for the opposite polarity. Such “forward” current flow can be approximated by the diode equation:

$$I(V) = I_o[\exp(V/nk_B T) - 1], \quad (1)$$

where I and V are the currents and voltages, respectively; n is the diode quality factor (a fitting parameter); k_B is the Boltzmann constant; and T is the temperature in Kelvin. The diode quality factor is often interpreted to account for “losses” or nonideal diode behavior. Some theories predict that the factor should fall between 1 and 2 depending on the recombination loss mechanism encountered. In practice, there is little correlation between solar cell performance and the actual n values used in fitting experimentally observed I - V characteristics. For moderately high-efficient cells, unpredicted values of $n > 2$ have been observed. The I_o factor depends on the height of the barrier, which in a homojunction cell is given by

$$I_o = I_{oo} * \exp(-E_g/mk_B T), \quad (2)$$

where m is a fitting parameter and I_{oo} a constant. Very often, the sum of two superimposed exponential curves is used to fit experimental $I(V)$ data. Theories have been developed to relate the magnitude of the fitting parameters to material properties, most commonly those of the absorber layer, which is often

assumed to be the most important factor controlling the performance of a solar cell.

When the diode is illuminated, a current flows in the reverse direction under the influence of the built-in field. If the device is short-circuited, the current density measured (per unit area of the cell surface) is called the short-circuit current density (J_{SC}). When the device is not connected to a load, the open-circuit voltage (V_{OC}) is measured. When the device is connected to a load, an I - V curve, as shown in Fig. 2 for a CuInGaSe₂-based solar cell, is obtained. Following the curve, a value for J and V can be found where the power output $J_{mp}V_{mp}$ is maximized (mp = maximum power). The ratio of $J_{mp}V_{mp}/V_{OC}J_{SC}$ is called the cell fill factor (FF). The power P generated by a solar cell is $P = FF \times V_{OC} \times J_{SC}$. The cell efficiency, η , relates the fraction of the total power of the incident sunlight (the solar constant, $S = 1 \text{ kW/m}^2$) to the electrical power generated by the cell (strictly speaking, under normal incidence and for a cell operating at a temperature of 20°C). Efficiency $\eta = P \text{ (kW/m}^2\text{)}/S \text{ (kW/m}^2\text{)}$ and is usually reported as a percentage. It is important to note that there can be a difference in the current-voltage behavior of the dark and the light current, commonly causing a “crossover” of the measured curves. Such observations clearly complicate the way in which solar cell efficiency can be related to the material properties used in a solar cell.

Another important property frequently measured is the solar cell collection efficiency and its spectral

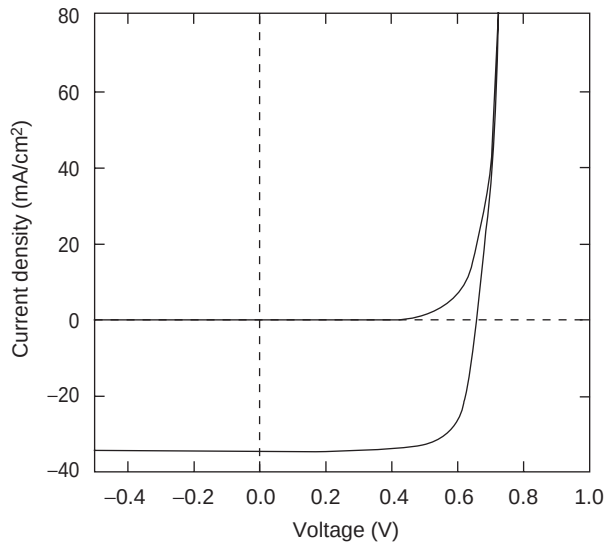


FIGURE 2 I - V curves under light and in the dark for a CuInGaSe₂ solar cell. Parameters are as follows: $V_{OC} = 0.66 \text{ V}$; $J_{SC} = 34.3 \text{ mA/cm}^2$; $FF = 77\%$; $\text{eff.} = 17.4\%$.

dependency. This is important because sunlight has a wide energetic distribution of photon energies. In the ideal case, all photons with energies larger than the bandgap E_g of the absorber generate one electron-hole pair, and ideally these can be separated and collected. Photons with energies $< E_g$ are not absorbed, thereby not creating carriers. Photons with energies much larger than E_g still only generate one electron-hole pair, and the excess energy is lost through thermalization because the carriers can be transported and collected only at the band edges. This fundamental limitation to solar cell efficiency is referred to as the Shockley–Queisser (SQ) limit after the authors who first considered this effect. Because of the SQ limitation, there is an optimum bandgap for the cell absorber; such an optimum is found by trading off the inability to absorb the low-energy solar photons in larger E_g absorbers and the need to increase the E_g and built-in potential (and thus voltage) of the cell. The optimum bandgap for a single absorber is approximately 1.4 eV . However, for E_g values between 1.0 and 1.8 eV , the dependency of the cell efficiency on the bandgap is not very strong, making all materials within this bandgap range candidates for solar cell absorbers. Because the electronic properties of a semiconductor may change as the bandgap is altered (e.g., by alloying), absorbers used in high-efficiency cells are often selected by making a compromise between bandgap and electronic performance.

A demonstrated way to overcome the SQ limitation is to stack two (or more) cells with different absorber bandgaps on top of each other. Thereby, the higher energy photons are absorbed in a higher bandgap top cell(s) (delivering higher voltage), whereas the low-energy photons pass through the top cell(s) and are converted in the lower E_g bottom cell(s).

Figure 3 shows the quantum efficiency (QE) of the cell from Fig. 2. The cell response is determined for light of single wavelength (or a narrow band of wavelengths), measuring the relative response with monochromatic or narrow-bandwidth light. Ideally, this curve would be almost a rectangle. In practice, optical and electrical collection losses determine the shape. There are also reflection, shadowing, and sometimes absorption losses as light enters the cell. It is often straightforward to account for such optical losses. It is more difficult to analyze the losses to the QE response that result from incomplete carrier collection. These losses may depend on the light penetration into the absorber—that is, how far away from the emitter (or collector, if the cell has a back-surface

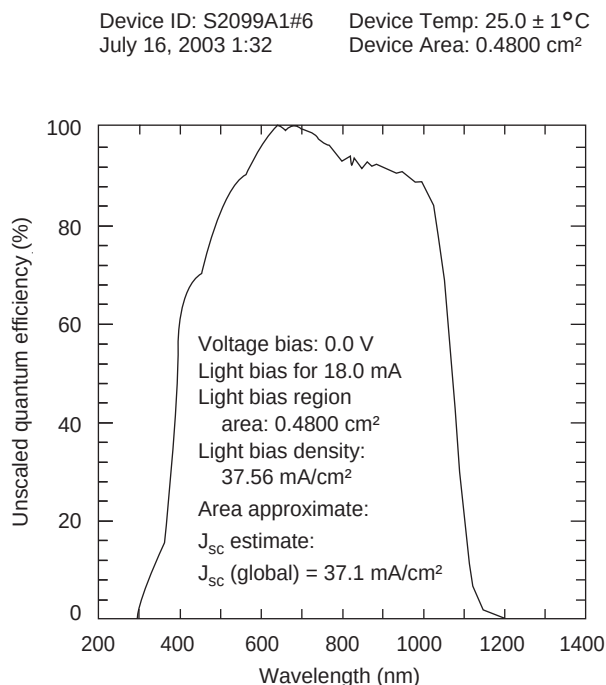


FIGURE 3 Spectral response for the cell shown in Fig. 2. The curve shows the relative response at each wavelength relative to the strongest response that defines the 100% value.

field) the carriers were generated. Except for amorphous and nanocrystalline thin-film silicon, organic, and dye-sensitized TiO_2 (Grätzel) solar cells, usually the absorber layers are thick enough to absorb most of the incident sunlight with photon energies $> E_g$. To first order, the shape of the long-wavelength fall-off of the QE is determined by the absorption edge of the absorber material. More detailed analyses reveal that the response depends on the details of cell design and the experimental conditions used, such as probe and bias light intensities, bias light wavelength, and electrical bias. For example, in crystalline Si solar cells, the edge of the long-wavelength QE response is determined by the doping of the wafer “base” (absorber). The base doping (at the levels typically used) does not change the optical properties of the Si wafer, but the response changes as if E_g were increased.

1.3 The Role of Buffer Layers

Solar cells are made from a wide variety of materials, and traditionally, less-than-ideal performance was attributed to very specific imperfections and defects found in the materials used to build the cells. It is also of interest to identify and characterize mechan-

isms and observations applicable to cells made from different materials. For many cell technologies, it has been a challenge for the experimental device fabricator to obtain cells with high V_{OC} . The opposite problem typically exists when a device structure is run in a computer model. The most important factor determining V_{OC} is the bandgap of the layers involved, so the initial modeling calculations often predict V_{OC} values greater than those observed experimentally. Although modeling may introduce ad hoc thin defective layers at the interface to match calculated and experimentally observed V_{OC} values, researchers usually maximize V_{OC} by employing grading schemes for the electronic and optical properties of the layers used near the junction. In thin-film solar cells, these schemes are typically referred to as buffer layers. In crystalline Si- and GaAs-based solar cells, these schemes are often referred to as contact passivation. It can be argued that a unique voltage-enhancing mechanism is operational in all solar cells built from different semiconductor materials.

Figure 4 presents a unifying buffer layer concept that can explain how buffer layers enhance solar cell voltages. Traditionally, it is assumed that the recombination loss current j_j would be determined by recombination through surface or interface defect states near the junction. Figure 4 suggests that, instead, the transport properties or carrier mobilities of the materials used to build the junction should be reduced (e.g., by using materials with potential fluctuations of the band edges that have significantly reduced carrier mobilities). This configuration reduces access of carriers to such voltage-reducing loss pathways. Solar cell fabrication depends on robust, repeatable, and scalable processes for creating an optimum or near-optimum grading of the transport properties in the materials used to build the junction regions of the cell. Ideally, the carrier-transport mobilities should be sufficient to allow carrier collection in the direction of the applied field but no larger. The principle can be best understood by comparing a Schottky barrier (SB; metal/semiconductor) crystalline Si junction and a metal/insulator/semiconductor (MIS) junction. Although working SB solar cells can be fabricated, they suffer from low cell voltage. On the other hand, MIS devices can attain voltages comparable to the best achievable by other junction-formation techniques. Although it is sometimes argued that the oxides provide passivation of the interface, it is well-known that for crystalline Si one or two atomic layers of thermally grown oxide provide excellent surface passivation. However, for

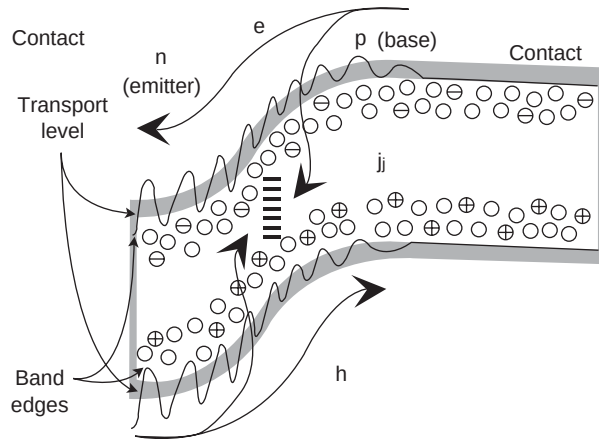


FIGURE 4 A unifying buffer-layer concept that can explain how buffer layers enhance solar cell voltages.

an MIS structure to exhibit much larger V_{OC} and device efficiency than the equivalent SB structure, the oxide buffer layer must be approximately 1.5 nm thick, nearly preventing carrier transport through this thin insulating layer. In a diffused crystalline Si p-n junction, the existence of a specific buffer layer is not generally acknowledged; however, a low-mobility buffer is likely present (i.e., a compensated layer as the phosphorus emitter diffuses into the boron-doped base). A buffer layer concept helps explain the trade-off between cell current and voltage that is often observed in the experimental research on optimizing solar cells. Because of its resistive nature, the buffer will cause a larger drop in the electric field near the junction, thereby reducing the extension of the space charge region in the absorber layer. Traditional cell analyses account for the measured device behavior in terms of surface and absorber recombination losses but cannot provide an explanation for the commonly occurring current-for-voltage trade-off observed in the optimization of experimental and commercial solar cells. It is important to realize that adding too much of a buffer layer can have disastrous consequences and ruin the solar cell performance. For industrial purposes, the thickness and electrical properties of the buffer layers must be uniformly and repeatedly obtained over large areas to achieve a consistent product.

2. CRYSTALLINE SILICON

2.1 Workhorse of the Industry

Most photovoltaic modules are made from crystalline silicon cells. Each cell is made from wafers produced

either from Czochralski single crystals or from large-grain polycrystalline ingots. Many photovoltaic manufacturers produce their own wafers in order to have tight control over manufacturing cost and wafer quality. In fact, in the early stages of technology development, inventions focused on deriving economical and performance advantages predominantly from the wafer-preparation process. The crystal growth processes appear to have become less critical in recent years because the structural perfection may be a lesser factor for cell performance than initially thought. Instead, much manufacturing cost reduction was derived from changing the wafering technology by switching from inner-diameter saws cutting one wafer at a time to wire saws slicing several feet of ingot in one cut. This has resulted in an increase in the number of usable wafers obtained per centimeter of ingot from 8–12 to 16–22, with typical wafer thicknesses at or slightly less than 300 μm and a further decreasing trend. In principle, wire sawing allows the cutting of much thinner wafers. However, it remains a challenge to process these thinner wafers and avoid excessive breakage during cell processing and module assembly.

Smaller, but not insignificant, portions of crystalline Si cells are based on wafers grown as silicon ribbon or as thick silicon films deposited on conducting ceramic substrates. These technologies avoid the wafering steps, and the performance of Si ribbon cells has reached levels comparable to those of wafer Si-based technologies. Of interest is that these various approaches still continue to be developed and used in parallel. Approximately 15 years ago, when today's manufacturers or their predecessors began to develop the different processes, each had hoped that their preferred method of making the Si substrate would result in a unique cost/performance advantage. It has not become evident which of these approaches is the most cost-effective. The reason for this may lie in the fact that the quality of the wafer plays a lesser role than originally thought. Instead, it became much more important to learn how to process each wafer into a high-efficiency solar cell.

2.2 Silicon Solar Cell Operation

For a crystalline Si solar cell, the complicated structure of Fig. 1 is often approximated in terms of a bulk (wafer) lifetime and a surface recombination velocity. Another frequently used parameter is the (zero-field) diffusion length as a measure for collecting carriers generated in the diffusion region of the absorber. Such characterization omits a crucial parameter: the cell voltage. For diffusion-processed cells,

V_{OC} depends critically on the boron doping level of the wafer base. Typical doping levels for optimum cell performance are 0.2–2 Ωcm . It is typically found that higher doping levels of the wafer or absorber increase V_{OC} , whereas they decrease J_{SC} . For the highest efficiencies, a compromise must be made.

Although most commercial cells are manufactured by diffusing emitters and collectors, higher performing champion Si solar cells often have a different cell structure. Point-contact or buried laser grooved-contact cells (referred to as passivated emitter, rear locally diffused cells) have reached efficiency levels of 24%. Because the localized contacts require large diffusion lengths to collect carriers over even greater distances, these high-efficiency cells can only be made with single-crystal “float zone” or Czochralski wafers. Many analyses have correlated V_{OC} values with the hole density in the absorber and have ascribed lower than optimum V_{OC} values to recombination losses in the absorber. With the realization of the so-called HIT (heterojunctions with intrinsic thin layers) solar cell developed by Sanyo Corporation, this interpretation becomes challenging. The HIT cell, a heterojunction Si cell, uses amorphous silicon emitter, collector, and buffer layers and n-type Czochralski wafers as the absorber. One practical advantage of this cell type is a temperature coefficient approximately half that in typical diffused cells, giving these cells an advantage at typical operating temperatures under full sunlight. The latter observations suggest that cell voltages and their temperature coefficients are predominantly determined by the details of junction formation and not by losses in the wafer or absorber.

The use of buffer layers in crystalline Si HIT cells supports the universality of a voltage-enhancing buffer layer concept. In diffusion-processed and screen-printed cells, fire-through nitride passivation layers are used. A nitride layer between the emitter and the metal grid contacts, frequently referred to as a passivation layer because of the high hydrogen contents of the nitride layers, can also be seen as a buffer concept. Finally, the sensitivity of V_{OC} to the boron base doping level in diffusion-processed cells can also be understood in terms of a buffer layer (in this instance, a low-mobility compensated layer that must be formed while the phosphorus emitter is diffused into the boron-doped p-type base).

2.3 Defect Engineering

Most crystalline Si cell manufacturers produce their cells by phosphorus diffusion for the emitter and use

aluminum alloy for the collector. It has become well accepted that the Si material can be upgraded during the cell-processing step. Figure 5 shows a pyramid schematic of how fundamental starting-material properties, substrate preparation, and processing during solar cell fabrication all strongly affect the resulting solar cell efficiency. Upgrading during cell processing often occurs automatically as the cell emitter, collector, and passivation and antireflection layers are formed, and it is usually referred to as phosphorous and/or aluminum gettering. Optimizing these gettering processes was critical in achieving today’s high commercial cell efficiency levels (typically 12 to >15%). Some manufacturers use additional process steps such as hydrogenation treatments for defect passivation. Such upgrades are commonly believed to lower the defect densities or the effectiveness of defects.

It is difficult *a priori* to distinguish whether the cell processes only affected the defect properties in the wafer (absorber region) or whether manipulation of the junction and buffer regions affected the space-charge region of the cell, improving the carrier collection from an unaltered absorber. The interactive nature of defects and impurities is well accepted among researchers. The lifetimes or diffusion lengths not only depend on the impurity levels (e.g., the concentration of a lifetime-killing metal impurity) but also depend, in a complicated way, on the presence of other inactive impurities, such as oxygen or carbon. Impurities interact with point or extended defects such as dislocations and structural defects in polycrystalline materials. The effectiveness of impurities in promoting recombination losses is strongly determined by

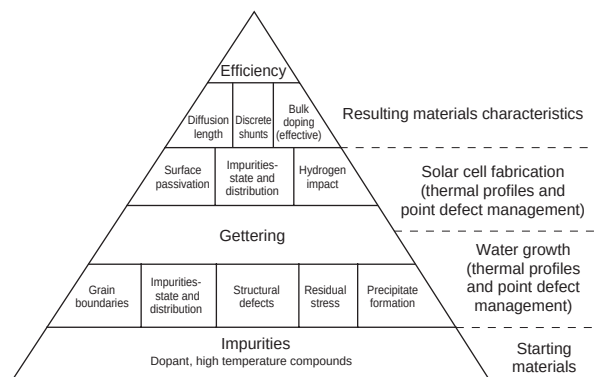


FIGURE 5 Pyramid schematic of how fundamental starting-material properties, substrate preparation, and processing during solar cell fabrication all strongly and interactively affect the resulting solar cell efficiency. Reproduced with permission from Rand *et al.* (2002).

gettering and passivation effects. In summary, it is difficult to describe solar cell performance simply in terms of defect and impurity concentrations in the Si wafer or sheet material alone.

3. THIN-FILM SILICON

3.1 Amorphous Silicon

Amorphous silicon solar cells are commercially available and can be produced on a variety of substrates ranging from glass to flexible thin foils. Cells are built in p-i-n or n-i-p configurations, where p and n represent thin doped (amorphous or nanocrystalline) layers, and the absorber layer is an intrinsic undoped layer. Cells are deliberately made thin (<300 nm thick) such that the space-charge region extends through the entire absorber (i-layer). The very small thickness is possible because amorphous silicon has stronger light absorption than crystalline Si (a direct bandgap) and light-trapping schemes have been developed to enhance the optical absorption.

The semiconducting properties of this material were discovered when films were prepared by plasma-enhanced vapor deposition of silane (SiH_4) gas. It was established that the optical and electrical properties are dominated by approximately 10 atomic-% (at.%) hydrogen in the film. It was also discovered that the material could be doped n-type (usually with phosphorus) and p-type (usually with boron), making this material much like a crystalline semiconductor. Until then, there was debate about whether amorphous semiconductors could, in principle, be doped. Some argued that the flexibility of the amorphous network would prevent three-valent acceptor or five-valent donor atoms from being incorporated into the four-valent Si network, which is necessary to produce additional conduction electrons or holes. Doping changes the resistivity of amorphous silicon by more than seven and six orders of magnitude for n- and p-type doping, respectively.

Unhydrogenated amorphous silicon films can also be prepared, for example, by evaporating silicon onto a substrate below 400°C . However, such films are very defective and show high conductivity. It is generally accepted that the quality of such films is low because the very high density ($>10^{20}/\text{cm}^3$) of silicon “dangling bonds” acts as a defect. With hydrogenation, their density is orders of magnitude lower (10^{15} – $10^{17}/\text{cm}^3$). Device-grade intrinsic hydrogenated amorphous silicon (a-Si:H) has the

following properties: bandgap, 1.6–1.8 eV; dark conductivity, 10^{-9} – $10^{-12} \Omega\text{cm}$; and photoconductivity under 1 sun intensity light of 10^{-5} – $10^{-6} \Omega\text{cm}$. The conductivity activation energy ranges from 0.65 to 0.9 eV.

Amorphous silicon degrades under illumination. In a solar cell, all three parameters— V_{OC} , FF, and J_{SC} —can be affected and usually decrease upon illumination. In intrinsic layers, the degradation affects the photoconductivity (which decreases), and the midgap density D_0 , which increases. The properties of the light-induced D_0 defect (e.g., its paramagnetic spin resonance signal) are virtually indistinguishable from those of the native D_0 defect found in the annealed state in much lower concentrations. Fortunately, the degradation appears to be self-limiting and stabilizes. The exact rates of degradation depend on the material properties and the conditions (temperature and light intensity) used for light soaking. Annealing (for 5–150 min at 150 – 180°C) essentially restores the light-soaked material (or cell) to its original state. With 1 sun intense light, most of the degradation will occur within several tens to hundreds of hours. This effect is referred to as the Staebler–Wronski mechanism.

It is clear that the effect, although influenced by the presence of impurities, is inherent to the amorphous silicon– (or germanium–) hydrogen alloy system. There is no sure mitigation scheme to avoid degradation in high-quality a-Si:H, but the details of materials preparation can noticeably influence the amount of material property or solar cell degradation observed. Because of the intrinsic nature of the effect, it is common to examine stabilized materials and solar cell performance or efficiency.

In the past, the near-midgap dangling bond defect was used as a strong quality indicator. Specifically, models were made that correlated such defect density with solar cell fill factors and the photoconductive properties of the layer. However, careful quantitative work later showed that the solar cell fill factor or the photoconductivity (PC) of the intrinsic layer are not controlled by a simple, single, straightforward mechanism. In the past, it was assumed to be the generation of D_0 defects. It has been confirmed that the kinetics with which observables such as D_0 , PC, or cell FF change upon light soaking or annealing show differences, and there is a lack of quantitative correlation between D_0 generation and PC and FF losses but a better correlation between PC and FF. Such findings may explain why materials with lower stabilized D_0 densities have not led to improved stabilized solar cell performance.

Also, degradation and annealing behavior in most instances can be classified into two mechanisms: fast (low-temperature annealable) and slow (below 80°C nonannealable) mechanisms. Any comprehensive Staebler–Wronski model must account for both mechanisms, which appear to be linked. The effects of the two mechanisms on degradation and annealing are large enough to affect the power output of a-Si:H-based modules by $\pm 10\%$. Interest in mitigating Staebler–Wronski degradation remains high because if it were possible to maintain today's solar cell performance in the higher annealed state, 16% efficient multijunction cells and 10% efficient commercial modules would likely be a reality. However, all data suggest that the degradation is intrinsic to a-Si:H and related alloys.

The prevailing mechanisms suggested in the literature implicate weak Si–Si bonds and weakly bonded hydrogen in the material as the cause for the instability. Experimentally, the instability appears not to depend systematically on hydrogen content of the material or on how tightly hydrogen is bonded. In fact, some of the more stable (deuterated and high hydrogen dilution, lower substrate temperature) solar cell recipes were achieved using materials that contain larger amounts of loosely bonded hydrogen.

Amorphous silicon solar cell technology has realized the first low-cost multijunction bandgap commercial solar cells. To make the cells more sensitive to red light, a-Si:H is alloyed with germanium (Ge). Especially for Ge contents > 50 at.% (corresponding to a bandgap of < 1.45 eV), the solar cell performance decreases significantly. A plausible explanation for this phenomenon is the decreased carrier collection due to lower carrier mobilities or higher recombination losses. This decreased carrier collection is mitigated by making the narrow bandgap cells very thin (150 nm or less) and by grading the Ge content in the intrinsic layers, such that the highest Ge content is found near the p-i interface of the junction. Use of Ge alloying for multi-bandgap cells has increased the stabilized device performance (e.g., on the order of 1 absolute efficiency point when an a-Si:H/a-Si:H cell and a-Si:H/a-Si_{0.5}Ge_{0.5}H cell are compared). However, Ge alloying has enhanced the cell performance less than expected from the reduction in bandgap because of deterioration of the electronic properties with Ge alloying and significant light-induced degradation in such alloys, which can be offset only by using a very thin, Ge-graded, narrow bandgap cell structure. Attempts were made to use carbon alloys (a-SiC:H) to widen the bandgap of the top cell of a

triple-junction device to increase the voltage of this component cell. Such efforts have been abandoned because a-SiC:H cells showed aggravated light-induced degradation, and it was not possible to increase the voltage of these cells beyond the best values achievable with optimized a-Si:H intrinsic layers (1.0 V). There is no consensus as to which microscopic mechanisms cause the electronic deterioration upon alloying with Ge, C, or, even more significant, other elements.

3.2 Crystalline Silicon Films

There are two approaches for using thin silicon films for photovoltaic applications. When amorphous silicon deposition techniques are used with excessive hydrogen dilution, the resulting films become nanocrystalline (nc-Si). The term microcrystalline has also been used for such films; however, grain sizes are always $\ll 1 \mu\text{m}$. Typical crystallite sizes are on the order of tens of nanometers, and to have reasonable electronic properties, a few atomic percent of hydrogen must be incorporated in such films. Solar cells with efficiencies of approximately 7% have been demonstrated using low-temperature deposition approaches.

Conventional thinking caused researchers to search for higher solar cell efficiency using Si films with much larger grain sizes. However, experimental work established a lack of improvement of cell efficiency with grain size in the range of 100 nm to 1 mm. To achieve grain sizes $> 1 \text{ mm}$, either high-temperature recrystallization (remelting) or epitaxial growth on a larger grain silicon substrate are required. Such methods require a substrate that can withstand the high temperatures involved and that is chemically inert so as not to release impurities into the silicon film. Such requirements may eliminate commercial glasses as a substrate. In the laboratory, graphite and ceramic substrates have been used with some success.

First, the low-temperature approaches are discussed. The solar cell efficiencies achieved appear too low to justify the development of a single-junction nanocrystalline solar cell technology. The main application is replacing the bottom cell in a-Si multijunction cells. To function as a bottom cell, the nc-Si cell must generate $> 24 \text{ mA/cm}^2$ and have a sufficiently high red response to generate $> 8 \text{ mA/cm}^2$ in a triple-junction stack cell or $> 12 \text{ mA/cm}^2$ for a double-junction device. With a high-quality back reflector for optical enhancements, current densities

of approximately 24 mA/cm^2 were demonstrated with nc-Si absorbers $\leq 2 \mu\text{m}$ thick.

How nanocrystalline cells operate is very poorly understood. Preparation of nc-Si films and cells is extremely sensitive to deposition conditions. Physical properties that may change are the crystalline orientation, luminescence properties, and conductivity of the films. When such changes occur, solar cell properties also change, often dramatically. However, it is difficult to pinpoint how the properties of the films determine cell performance. Rather, whenever the absorber properties change, it is necessary to readjust the junctions (p, n, and buffer layers) used in the cell. It appears from the optimization of nc-Si cells that a regime is often encountered in which higher J_{SC} values are obtained by sacrificing V_{OC} or vice versa. Such a trade-off between voltage and current is not uncommon for both thin-film and wafer-based crystalline silicon. The highest voltages for the nc-Si cell, in conjunction with the required high J_{SC} values, remain to be determined. Better optimization and better understanding of the mechanism leading to such a trade-off are both required.

Cells spanning the amorphous to nanocrystalline transition (adding a gradually increasing nanocrystalline phase to the intrinsic layer) have been prepared. Cells with a large fraction of amorphous phase in their i-layer show large voltages but also light-induced degradation. Light soaking actually enhances V_{OC} but decreases FF. With such unstable i-layers, the high J_{SC} values required for multijunctions could not be demonstrated. It is important to note that even the stable cell recipe with the largest grain size and best-ordered i-layers did not result in the highest efficiency.

Two additional universal findings about nc-Si cells should be mentioned. First, as for a-Si:H alloys, the best cell results are obtained when the i-layer deposition rate is slow ($< 0.5 \text{ nm/s}$). However, some deposition methods [such as hot-wire chemical vapor deposition (CVD)] have shown that different film structures and modest cell efficiencies (routinely, approximately 5%) can be obtained at higher deposition rates ($1\text{--}2 \text{ nm/s}$). Second, it has been observed that some nc-Si films are porous and therefore subject to postdeposition oxidation. After oxidation, solar cell performance degrades severely. Such films must be avoided because it has been reported that oxidation occurs even when such devices are capped, for example, by a p-layer in a solar cell.

Of concern with regard to nc-Si is that in recent years cell efficiency records have been somewhat stagnant (approximately 10%). Given the current

performance, it appears that there is opportunity to incrementally improve a-Si:H-based multijunctions by using the nc-Si cell as a bottom cell. Given the large variety of nc-Si films that can be produced, it may well be possible that cell recipes, at greater deposition rates, can be found that result in a breakthrough for even higher cell efficiency.

Next, the higher temperature, large-grain crystalline Si films are discussed. Two approaches have resulted in larger grain sizes. One approach to thin-film solar cells relies on melt recrystallizing a CVD-deposited, $\sim 100\text{-}\mu\text{m}$ -thick growth layer on a ceramic or other substrate. Then, active solar cell layers are deposited epitaxially by CVD. The active Si layers used are a $5\text{-}\mu\text{m}$ -thick back-surface field and contact layer (boron-doped, $\sim 0.1 \Omega \text{ cm}$) and a very lightly boron-doped, 30- to $35\text{-}\mu\text{m}$ -thick absorber layer. The cell emitter is formed by a P-diffusion process, resulting in a cell with the following parameters: $\eta = 9.1\%$, $V_{\text{OC}} = 0.54 \text{ V}$, $\text{FF} = 0.73$, and $J_{\text{SC}} = 23.1 \text{ mA/cm}^2$. Because the ceramic substrate is insulating, monolithic interconnection of cells can be accomplished and has been demonstrated. Interestingly, the current generated and collected in these absorbers is similar to the current densities measured in the 1.5- to $2\text{-}\mu\text{m}$ -thick nc-Si absorbers. This may be because in thin, optically enhanced Si film cells, carrier collection is quite efficient, whereas in thicker films carrier collection remains a challenge for carriers generated some distance (more than a few micrometers) from the junction.

Other deposition approaches, such as using an iodine vapor transport reaction, provide a process for growing large-grain films in the $750\text{--}950^\circ\text{C}$ substrate temperature range. Cell efficiencies of these larger grain silicon films have not met expectations. One reason may be that grain size is not as much a driver of material quality and cell performance as has been hitherto assumed. The behavior of the nanocrystalline cells (in which smaller grain size leads to higher cell performance) suggests that there may be a genuinely beneficial aspect of polycrystallinity. For intermediate grain size silicon films, it may be that this effect is lost, and the carrier collection is still too impeded compared to that of single- or large-grain ($> 1 \text{ mm}$) crystalline Si wafers.

4. POLYCRYSTALLINE THIN FILMS

Since approximately 1990, solar cell efficiencies of $> 13\%$ have been obtained with copper indium diselenide (CuInSe_2 or CIS) and cadmium telluride

(CdTe) polycrystalline thin-film semiconductors. These materials have few other applications. They are attractive as solar cell absorbers because they perform well. Cell efficiencies on the order of 19% for CIS and 16% for CdTe have been demonstrated. These materials were mostly developed using heuristic solar cell optimization schemes, with a lagging understanding of the fundamental physical properties of the materials and cells.

4.1 Copper Indium Diselenide

CIS was developed after it was realized that CdS/CuS solar cells developed earlier, despite promising cell efficiency (10% demonstrated), were not viable because of their propensity to become shunted while operating. This shunting was related to copper migration. On the other hand, the ternary CuInSe_2 and quaternary $\text{CuIn}_{1-x}\text{Ga}_x\text{Se}_2$ compounds (x is typically 0.2–0.3 for high-efficiency cells) do not suffer from such intrinsic degradation. Early on, it was realized that to obtain high solar cell performance, the films had to be nonstoichiometric. Specifically, for good cell performance, the $\text{Cu}/(\text{In} + \text{Ga})$ atomic ratio must be between 0.86 and 0.96. The Se content of the film must be slightly higher than 2. Films for solar cells are deposited by a variety of methods, including coevaporation from metal or metal–selenide precursors. It is beneficial to have some excess selenium available during deposition. However, it is also possible to react mixed-phase metal or metal–selenide precursors into a single-phase (or seemingly single-phase) CuInGaSe_2 layer in a selenization reaction in the presence of Se vapor or H_2Se gas. Another approach is to produce absorbers from nanoparticle precursors. Many deposition recipes for high-efficiency cells call for ramping both the substrate temperature and the ratios of the metal or metal–selenide precursors during film growth such that the Cu deficiency is greatest near the surface of the film. The addition of Ga has made it easier to achieve higher efficiency cells, both because the bandgap of the absorber increases from approximately 1.0 to 1.15 eV and because Ga affects the electronic properties in a way that is favorable to cell performance, especially the V_{OC} parameter. For the best cell results, Ga incorporation in the absorber should be uniform or in a “notch” distribution (higher concentrations near the front and back of the absorber). The best solar cell results are obtained for films deposited or selenized at temperatures higher than 500°C. Unfortunately, some deposition approaches lead to an

undesirable high accumulation of Ga mostly near the back surface of the absorber. In these instances, sulfur has been used to form $\text{Cu}(\text{In,Ga})\text{Se}_{2-x}\text{S}_x$ pentenary compounds to achieve the bandgap widening and desirable electronic modification near the front surface of the absorber. However, experimental results suggest that S alloying leads to somewhat inferior cell performance compared to optimized Ga alloying.

CIS solar cells are built up on a molybdenum (Mo)-coated soda-lime glass substrate. The use of Mo metallization is important to achieve high cell performance, but the reason for this is poorly understood. It is well established that the addition of some sodium during film growth is quite beneficial; Na had been incorporated inadvertently when soda-lime glass substrates were used. Sodium can be added during absorber growth to achieve high performance on Na-free glasses or on Mo-coated metal or high-temperature resistant plastic foils.

CIS devices use a ZnO/CdS heterojunction emitter layer. The best and most consistent solar cell results are obtained when a thin CdS layer is deposited using a wet chemical bath deposition (CBD) process. The reason for this is debated in the literature, and no consensus has been reached. Alternatives to CdS (such as resistive oxides or ZnS) have been explored with some success but have not achieved the same champion cell efficiencies or processing robustness as CBD CdS. The ZnO contact is often prepared in the form of a graded layer or bilayer, where the part of the film in contact with the CdS is deliberately made more resistive. It can be argued that in a CIS solar cell the optimum junction profile is affected by the resistive ZnO layer, the CdS or alternative buffer layer, and the surface portion of the $\text{Cu}(\text{In,Ga})\text{Se}_2$ absorber. This makes the cell optimization process very interdependent; that is, if one of the layers in the cell is altered, all other layers (their properties and thicknesses) must be reoptimized. In other words, there is no unique optimum for the properties of each layer. Conventionally, it is assumed that cells fabricated from more perfect layers result in higher cell efficiency. For CIS and CdTe solar cells, cells deposited as multicrystalline thin-film layers on low-cost substrates have achieved greater device performance than solar cells made on (bulk or epitaxial) single-crystal versions of these materials.

4.2 CdTe Solar Cells

CdTe was developed after it was realized that a number of deposition approaches, such as

evaporation, sublimation, sputtering, or electrodeposition, resulted in promising devices (10% demonstrated). Early on, researchers realized that to obtain high solar cell performance, the films had to be heat treated after exposing them to a CdCl_2 solution or vapor. Initially, growth of the crystallite grain size and Cl incorporation into the film were observed. However, it was later found that the CdCl_2 heat treatment is also essential for cells with CdTe absorber layers grown at high temperature having even larger grain sizes. Treatments are carried out in the range of 360–500°C. CdTe always grows as a stoichiometric compound that is much less toxic than Cd metal. There is debate as to whether oxygen incorporation (doping) of the CdTe layer is necessary or beneficial.

CdTe cells usually use a SnO_2 /buffer/CdS heterojunction emitter layer. It is important to keep the thickness of the CdS as thin as possible to avoid optical absorption losses from this layer. However, reducing the thickness too much often leads to reduced V_{OC} and FF. The buffer layers are often undoped SnO_2 layers. Some alternative buffer layers, such as zinc stannate or CdS:O (the latter prepared as an amorphous layer that crystallizes during the cell anneal step), may ultimately lead to higher cell performance. It is generally accepted that some intermixing of the CdS and CdTe takes place and that such intermixing may positively affect cell performance.

The rear-contact preparation for CdTe is more an art than a science. Unsuitable contacts result in very poor device performance. At first, for laboratory cells, a bilayer of copper and gold was used with good success. All contact applications generally benefit from acid etching the CdTe device. This has been shown to leave a Te-enriched surface. The Cu/Au contacts, in general, show poor long-term stability and are not suitable for commercial product. Graphite pastes were found to be another good contacting scheme; graphite paste (“dac”) that is typically doped with Cu has been applied to the etched surface. Usually, conductive silver ink is applied as a secondary contact to the graphite layer. This type of contact performs well but is less desirable as a contacting scheme for large-area commercial modules. Because contact formation was difficult due to the nearly intrinsic nature of the CdTe layer, Cu-doped ZnTe layers have also been successfully developed as a back contact. Although they are thought of as completing the junction, it has been reported that the choice of metal used to contact this ZnTe layer can affect the solar cell

properties. Commercial manufacturers have developed proprietary thin-film contacting schemes.

It is remarkable that the performance of a CdTe solar cell is very strongly affected and can be dominated by either junction (i.e., the emitter or the collector). Early commercial CdTe modules have shown some change (sometimes an initial improvement, but also degradation) in performance. Laboratory cells are typically stressed at 60–90°C and show some loss of power (0–30%) after such stressing, unless the cells are processed in a very narrow, empirically established process window. The stability is affected by the details of both the front and back contact as well as by the CdCl_2 anneal treatment of the CdTe. The rapid degradation of cells using Cu/Au contacts suggests Cu diffusion or migration as a major cause. Further studies have revealed, however, that although too much Cu applied with the back contact leads to degradation, variations for modest amounts of Cu (e.g., in Cu-doped graphite or Cu-doped ZnTe contacts) can increase or decrease cell stability. Cells made without any Cu deliberately added were also found to degrade. Thus, minimizing or eliminating the Cu doping of the back contact is not a sure way to eliminate the degradation. It appears that cells and modules with sufficient commercial stability (<10–20% degradation over 20 years of operation) can be manufactured, but each manufacturing process must be tested to eliminate processes that result in too much degradation.

5. CONCENTRATOR SCHEMES AND III–V SOLAR CELLS

For large centralized photovoltaic arrays, it has been proposed to minimize the use of semiconductor material by focusing the sunlight through lens or mirror systems onto a solar cell mounted at the focal point of the concentrator. Such systems must be “tracked” (i.e., rotated around two axes to remain in focus with the sun). This approach opens up the possibility of using very high-efficiency solar cells that would be too costly to incorporate into flat-plate arrays for terrestrial power production. Because of the complexity and maintenance requirements of mechanical trackers, this approach is only viable for large systems in which hundreds of kilowatts or perhaps megawatts of power are required. An advantage of this scheme is that the efficiency of high-efficiency solar cells increases further under concentrated sunlight. A drawback is that such

systems have some optical and system losses and produce power only from direct sunlight (i.e., when the sun is directly visible), whereas flat-plate systems still convert the 5–50% of sunlight available during cloudy conditions when the sun is partially or totally obscured.

The solar cells for such concentrator systems are either high-efficiency single-crystal silicon solar cells (20–27% conversion efficiency under concentration) or GaAs-based multijunction solar cells that have demonstrated efficiencies well above 30% under concentration. Champion Si cells and GaAs cells have attained close to theoretically possible efficiency values, and the challenge is to incorporate such cells into even higher efficiency multijunction solar cells.

The GaAs or III–V semiconductor community has developed near-ideal high-efficiency multijunction solar cells. Typically, these cells are built as homo-junctions. For high performance, combinations of a lower doped n-type homoemitter and a higher doped (AlGaAs) heteropassivation layer are used. Without concentration, these high-efficiency cells are also used commercially for space cells, for which their high efficiency and radiation resistance outweigh their much higher cost. The cells are grown in thin-film form using heteroepitaxy on a Ge substrate. A popular cell structure is the GaInP/GaAs/Ge triple-junction cell, providing the appropriate lattice-constant match to Ge and GaAs for successful heteroepitaxy growth. In principle, although for the III–V semiconductors there are a wide variety of compounds providing a very large range of E_g (while still being lattice matched to Ge and GaAs), some materials clearly work better than others for solar cells. There has been no success in achieving high cell performance for a 1-eV bandgap cell using lattice-matched $\text{Ga}_{1-x}\text{In}_x\text{As}_{1-y}\text{N}_y$ alloys. Even for this near-ideal and well-characterized semiconductor system, it is unknown which physical/chemical absorber properties determine high and low solar cell performance.

6. EMERGING SCHEMES

6.1 Dye-Sensitized

A solar cell efficiency greater than 10% has been reported by Michael Grätzel *et al.*, who invented and optimized a dye-sensitized solar cell at the Swiss Federal Institute of Technology. The dye-sensitized cell is an electrochemical cell in which an electrolyte, rather than a conductor or semiconductor, is used to make one of the junctions of a solar cell. The

absorber layer consists of a deliberately porous SnO_2 /titanium dioxide (TiO_2) coating, for which the surface of the TiO_2 is on the order of 1000-fold enhanced over the surface of a smooth film. This ~ 100 -nm-thick TiO_2 is coated with a very thin (~ 1 -nm-thick) dye. The best performance has been achieved with ruthenium (Ru)-containing dyes, such as Ru bipyridyl or Ru terpyridine complexes that absorb near-infrared light. The contact is made with an iodine redox electrolyte and a second piece of SnO_2 -coated glass as the electrical contact. Note that this type of cell is not actually wet. The electrolyte simply has to wet the dye, requiring only drops of electrolyte per 100-cm^2 device area. Nevertheless, when multiple cells are made on one substrate with monolithic interconnection, this device must be tightly sealed not only around the perimeter but also between adjacent cells. In an attempt to simplify fabrication, solid or gel electrolytes have been investigated but have resulted in inferior cell performance compared to cells with liquid electrolytes.

The prevailing physical description of cell operation assumes that in this device the photogenerated electrons and holes are already separated upon generation in the physical phase boundary of the dye. This creates a photo-induced potential gradient that drives electrons and holes in opposite directions.

6.2 Other Schemes

Although semiconducting organic polymers have been known for some time, only recently has there been a focus on using these materials for preparing solar cells. Their physics and limitations are poorly understood. The junction is made from an electron conductor (e.g., perylene benzimidazole) and a hole conductor (e.g., poly-3-hexylthiophene). It is generally believed that in these materials the absorbed light creates excitons that are transported to the planar interface at which the carriers are separated. The efficiency levels achieved (up to 4% in a few instances but typically lower) suggest that further development is required before these will become candidates for terrestrial power application. Interest is high because in recent years there have been significant advances with regard to organic light-emitting diodes in terms of luminosity.

6.3 Theoretical Concepts

Several schemes have been proposed to overcome the SQ limit discussed previously. Aside from multijunction solar cells, which have been produced and

allowed moderate efficiency gains, schemes have also been proposed in which suitable arrangements in single-junction devices may lead to cell efficiencies exceeding 80%. One proposal suggests that sub-bandgap light can be absorbed via defect states (impurity solar cells). Other schemes propose using the energy that is greater than the corresponding bandgap of the absorber to create additional electron hole pairs (hot carrier cells). These schemes rely on separating the generation and collection processes of carriers, and they assume lossless carrier collection. Experimentally, attempts to realize such schemes have been unable to verify the expected significant enhancement of the short-circuit current density or spectral response.

7. FUTURE POTENTIAL (PHYSICAL LIMITATIONS OF MATERIALS)

There is good empirical knowledge that some materials work better than others for solar cells. In the search for new materials, materials with the desired bandgap, electronic properties allowing carrier separation and collection, and electronic properties that allow robust junction-fabrication schemes must be found. Researchers with a background in GaAs-based semiconductors have also stressed the importance of matching the lattice constants to achieve perfect epitaxial growth.

Achieving the desired bandgap (e.g., by alloying different materials) is possible most of the time. Cell efficiencies hinge on the electronic properties, which are often not conducive for achieving high cell performance. Often, it is difficult to determine what causes low cell efficiency. Structural perfection and lattice matching are requirements for GaAs-based solar cells. However, in other materials systems

disordered polycrystalline semiconductor films lead to solar cell performance equal to or exceeding that which can be achieved from single crystalline versions of these materials.

Robust junction-formation processes remain a challenge, although nature has provided some schemes for either homo- or heterojunctions that are quite reproducible, controllable, and manufacturable. It is becoming increasingly accepted that solar cell optimization is a very interdependent process; that is, if one layer of the cell is altered, leading to lower solar cell performance, this can be fixed by reoptimizing the processing of other layers in the cell. In other words, cell optimization cannot rely on combining individually optimized layers to achieve high performance. Rather, the correct combination of absorber, junction, and contacting layers is required to achieve high solar cell performance.

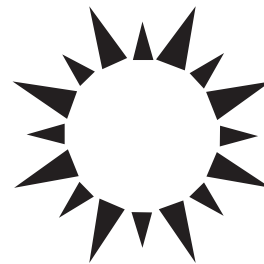
SEE ALSO THE FOLLOWING ARTICLES

Materials for Solar Energy • Photovoltaic Conversion: Space Applications • Photovoltaic Energy: Stand-Alone and Grid-Connected Systems • Photovoltaics, Environmental Impact of

Further Reading

- Luque, A., and Hegedus, S. (2002). "Handbook of Photovoltaic Science and Engineering." Wiley, New York.
- Rand, J., Rozgonyi, G., Lu, J., and Reedy, R. (2002). Characterization of silicon-film sheet material. Conference Record of the 29th IEEE Photovoltaic Specialists Conference, New Orleans, May 19–24, pp. 98–101.
- Shockley, W., Queisser, H. J. (1961). Detailed balance limit of efficiency of p–n junction solar cells. *J. Appl. Phys.* **32**, 510–519.

Photovoltaics, Environmental Impact of



VASILIS M. FTHENAKIS

Brookhaven National Laboratory
Upton, New York, United States

ALEC O. BULAWKA

U.S. Department of Energy
Washington, D.C., United States

-
1. Overview of Hazards in Photovoltaic Manufacture
 2. Crystalline Silicon Solar Cells
 3. Amorphous Silicon Solar Cells
 4. Cadmium Telluride Solar Cells
 5. Copper Indium Diselenide Solar Cells
 6. Gallium Arsenide High-Efficiency Solar Cells
 7. Operation of PV Modules
 8. Decommissioning PV Modules
 9. Environmental Benefits of Photovoltaics Operation
 10. Conclusion

Glossary

emergency response planning guideline-2 (ERPG-2) The concentration of a hazardous material below which nearly all people could be exposed for up to 1 h without irreversible or other serious health effects or symptoms that would impair their ability to take protective action.

immediately dangerous to life or health concentration (IDLH) The maximum concentration of a hazardous material from which one could escape within 30 min without any symptoms impairing escape or any irreversible health effects.

threshold limit value—ceiling The concentration of a hazardous material that should not be exceeded during any part of the working exposure.

threshold limit value, short-term exposure level (STEL) The maximum concentration of a hazardous material to which workers can be exposed for up to 15 min, provided not more than four excursions per day are permitted with at least 60 min between exposures, and provided that the daily permissible exposure limit is also not exceeded.

threshold limit value, time-weighted average (TLV-TWA) The time-weighted average threshold concentra-

tion of a hazardous material above which workers must not be exposed during work shifts (8 h/day, 40 h/week).

Photovoltaic (PV) technologies have distinct environmental advantages for generating electricity over conventional power technologies. The operation of PV systems does not produce any noise, toxic emissions, or greenhouse gases; regardless of the composition or technology used, PV generation of electricity is a zero-emissions process. Photovoltaic modules are durable and do not produce emissions even during extreme conditions (accelerated aging with low Earth orbit radiation and thermal cycles from 80 to -80°C). Photovoltaic energy can help meet the growing worldwide demand for electricity and can do so without incurring the high economic and environmental costs of burning fossil fuels and installing power lines. Compared to burning coal, every gigawatt-hour of electricity generated by PVs would prevent the release of approximately 1000 tons of carbon dioxide, 10 tons of sulfur dioxide, 6 tons of nitrogen oxides, and 0.4 tons of particulates.

However, as is the case with any energy source or product, there are also environmental, health, and safety (EHS) hazards associated with the manufacture of solar cells. The PV industry uses toxic and flammable substances, although in smaller quantities than do most other industries, and using hazardous chemicals can entail occupational and environmental hazards. Addressing EHS concerns is the focus of numerous studies of the National Photovoltaic EHS Assistance Center at Brookhaven National Laboratory, which operates under the auspices of the U.S. Department of Energy. These studies are proactive and ongoing as new technologies, materials, processes,

and increased production levels are encountered. The activities of the center are coordinated with the U.S. PV industry, which is very sensitive and responsive to EHS issues. This article provides a concise summary of EHS issues pertaining to the manufacture of crystalline silicon, amorphous silicon, copper indium diselenide, copper indium gallium diselenide, gallium arsenide, and cadmium telluride solar cells, all of which are commercially available.

1. OVERVIEW OF HAZARDS IN PHOTOVOLTAIC MANUFACTURE

A variety of materials are used in manufacturing photovoltaic (PV) cells, including materials widely used in commercial applications (e.g., aluminum and glass), chemicals that have seen extensive use in the semiconductor industry (e.g., silane and hydrofluoric acid), and some chemicals that are unique to the PV industry [e.g., copper indium selenide (CIS) and cadmium telluride (CdTe)]. For such new materials, the U.S. Department of Energy has collaborated with the National Institutes of Health to study their potential health hazards. Table I presents the main hazards associated with specific technologies, and Table II shows that these hazards differ for different technologies and deposition processes. Potential risks are related to the materials' toxicity, corrosivity, flammability, and explosiveness.

2. CRYSTALLINE SILICON SOLAR CELLS

2.1 Occupational Health Issues

In manufacturing wafer-based crystalline silicon (*x*-Si) solar cells, there are some occupational health issues, including potential chemical burns and inhalation of fumes from the hydrofluoric acid, nitric

TABLE I

Major Hazards in PV Manufacturing

Module type	Potential hazards
<i>x</i> -Si	HF acid burns, SiH ₄ fires/explosions, Pb solder/module disposal
a-Si	SiH ₄ fires/explosions
CdTe	Cd toxicity, carcinogenicity, module disposal
CIS, CGS	H ₂ Se toxicity, module disposal
GaAs	AsH ₃ toxicity, As carcinogenicity, H ₂ flammability, module disposal

acid (e.g., HNO₃), and alkalis (e.g., NaOH) that are used for cleaning the wafers, removing dopant oxides, and cleaning the reactor. Dopant gases and vapors [e.g., phosphorus oxychloride (POCl₃) and diborane B₂H₆] are also hazardous if inhaled. POCl₃ is a liquid, but in a deposition chamber it can generate toxic P₂O₅ and Cl₂ gaseous effluents. Inhalation hazards are controlled with properly designed ventilation systems in the process stations. Other occupational hazards are related to the flammability of silane (SiH₄) and its by-products used in silicon nitride deposition.

2.2 Public Health and Environmental Issues

No public health issues were readily identifiable with this technology. The environmental issues are related to generating liquid and solid wastes during the slicing, cleaning, and etching of wafers and during the processing and assembling of solar cells.

The *x*-Si PV industry has initiated programs of waste minimization and hence studies environmentally friendlier alternatives for solders, slurries, and solvents. Success was reported at both laboratory and manufacturing scales in reducing the caustic waste generated by etching. Other efforts for waste minimization include recycling stainless-steel cutting wires, recovering the SiC in the slurry, and in-house neutralization of acid and alkali solutions.

Finally, the content of Pb in solder in many of today's modules creates concerns about the disposal of modules at decommissioning. One U.S. company is using Pb-free soldering and has offered assistance to the rest of the industry for adopting this option. It is likely that the *x*-Si industry will gradually move away from Pb-Sn solder.

3. AMORPHOUS SILICON SOLAR CELLS

Amorphous silicon (*a*-Si), CdTe, CIS, and gallium arsenide (GaAs) are thin-film technologies that use approximately 1/100 of the costly PV material used on *x*-Si.

3.1 Occupational Safety Issues

The principal safety hazard of this technology involves the use of SiH₄ gas, which is extremely pyrophoric. The lower limit for its spontaneous

TABLE II

Hazardous Materials Used in PV Manufacturing^a

Material	Source	TLV-TWA ^b	STEL ^c	IDLH ^d	ERPG-2 ^e	Critical effects
Arsine	GaAs CVD	0.05 ppm	—	3 ppm	0.5 ppm	Blood, kidney
Arsenic compounds	GaAs	0.01 mg/m ³	—	—	—	Cancer, lung
Cadmium compounds	CdTe and CdS deposition	0.01 mg/				m ³ (dust) 0.002 mg/
—	CdCl ₂ treatment	—	NA			m ³ (fumes)
Carbon tetrachloride	Etchant	5 ppm	10 ppm		100 ppm	Liver, cancer, greenhouse gas
Chlorosilanes	a-Si and α -Si deposition	5 ppm	—	800 ppm	—	Irritant
Copper	CIS deposition	1 mg/				m ³ (dust) 0.2 mg/
—	100 mg/m ³	—				m ³ (fumes)
Diborane	a-Si dopant	0.1 ppm	—	40 ppm	1 ppm	CNS, pulmonary
Germane	a-Si dopant	0.2 ppm	—	—	—	Blood, CNS, kidney
Hydrogen	a-Si deposition	—	—	—	—	Fire hazard
Hydrogen fluoride	Etchant	—	C ^f , 3 ppm	30 ppm	20 ppm	Irritant, burns, bone, teeth
Hydrogen selenide	CIS sputtering	0.05 ppm	—	1 ppm	—	Irritant, GI, flammable
Hydrogen sulfide	CIS sputtering	10 ppm	15 ppm	100 ppm	30 ppm	Irritant, CNS, flammable
Indium compounds	CIS deposition	0.1 mg/m ³	—	—	—	Pulmonary, bone, GI
Lead	Soldering	0.05 mg/m ³	—	—	—	CNS, GI, blood, kidney, reproductive
Nitric acid	Wafer cleaning	2 ppm	4 ppm	25 ppm	—	Irritant, corrosive
Phosphine	a-Si dopant	0.3 ppm	1 ppm	50 ppm	0.5 ppm	Irritant, CNS, GI, flammable
Phosphorous oxychloride	α -Si dopant	0.1 ppm				Irritant, kidney
Selenium compounds	CIS deposition	0.2 mg/m ³		1 mg/m ³		Irritant
Sodium hydroxide	Wafer cleaning		C ^f , 2 mg/m ³	10 mg/m ³	5 mg/m ³	Irritant
Silane	a-Si deposition	5 ppm	—	—	—	Irritant, fire and explosion hazard
Silicon tetrafluoride	a-Si deposition	—				
Tellurium compounds	CIS deposition	0.1 mg/m ³				CNS, cyanosis, liver

^a Abbreviations used: CVD, chemical vapor deposit; NA, not available; CNS, central nervous system; GI, gastrointestinal tract.^b The threshold limit value, time-weighted average (TLV-TWA) is defined by the American Conference of Governmental Industrial Hygienists (ACGIH) as the time-weighted average threshold concentration above which workers must not be exposed during work shifts (8 h/day, 40 h/week).^c The threshold limit value, short-term exposure level (STEL) is defined by the ACGIH as the maximum concentration to which workers can be exposed for up to 15 min, provided not more than four excursions per day are permitted with at least 60 min between exposures, and provided that the daily permissible exposure limit is also not exceeded.^d The immediately dangerous to life or health concentration (IDLH) is defined by the National Institute for Occupational Safety and Health (NIOSH) as the maximum concentration from which one could escape within 30 min without any symptoms impairing escape or any irreversible health effects.^e The emergency response planning guideline-2 (ERPG-2) is defined by the American Industrial Hygiene Association as the concentration below which nearly all people could be exposed for up to 1 h without irreversible or other serious health effects or symptoms that would impair their ability to take protective action.^f The threshold limit value—ceiling (C) is defined by the ACGIH and the NIOSH as the concentration that should not be exceeded during any part of the working exposure.

ignition in air is 2 or 3%, depending on the carrier gas. If mixing is incomplete, a pyrophoric concentration of SiH_4 may exist locally, even if its concentration in the carrier gas is less than 2%. At silane concentrations $\geq 4.5\%$, the mixtures were found to be metastable and ignited after a certain delay. In an accident, this occurrence could be extremely destructive because the protection provided by venting would be ineffective. In addition to SiH_4 , hydrogen used in a-Si manufacturing is flammable and explosive. Most PV manufacturers use sophisticated gas-handling systems with sufficient safety features to minimize the risks of fire and explosion. Some facilities store silane and hydrogen in bulk from tube trailers to avoid the need to frequently change gas cylinders. A bulk module typically contains eight cylindrical tubes that are mounted onto a trailer suitable for short- and long-range transport. These modules carry up to 3000 kg of silane. Another option utilizes a single, 450-liter cylinder mounted on a skid, which contains up to 150 kg of silane (minibulk). For additional safety, these storage systems are equipped with isolation and flow-restricting valves.

Bulk storage decreases the probability of an accident since changing the trailer is an infrequent event that is executed in a precise, well-controlled manner under the supervision of the plant's management, safety officials, the gas supplier, and local fire department officials. On the other hand, if an accident does occur, the consequences can be much greater compared to those involving gas cylinders. Silane is used mainly in glow discharge deposition at very low utilization rates (e.g., 10%). To the extent that this rate increases in the future, the potential worst consequences of an accident will be reduced.

Toxic doping gases (e.g., AsH_3 , PH_3 , and GeH_4) are used in quantities too small to pose any significant hazards to public health or to the environment. However, leakage of these gases can cause significant occupational risks, and management must provide continuous vigilance to safeguard personnel.

3.2 Public Health and Environmental Issues

Silane, used in bulk quantities in a-Si facilities, may pose hazards to the surrounding community if separation zones are inadequate. In the United States, the guidelines of the Compressed Gas Association specify minimum distances from places of public assembly that range from 80 to 450 ft, depending on

the quantity and pressure of silane in the containers used. The corresponding minimum distances to the plant's property lines are 50–300 ft. These prescribed separation distances are considered sufficient to protect the public under worst-condition accidents. No environmental issues have been identified with this technology.

4. CADMIUM TELLURIDE SOLAR CELLS

4.1 Occupational Health Issues

In CdTe manufacturing, the principal concerns are related to the toxicity of the feedstock materials (e.g., CdTe, CdS, and CdCl_2). The occupational health hazards presented by Cd and Te compounds in various processing steps vary as a function of their specific toxicity, physical state, and the mode of exposure. No clinical data are available on the human health effects of exposure to CdTe. The National Institute of Environmental Sciences conducted a limited number of laboratory tests comparing the acute toxicity of CdTe, CIS, and copper indium gallium diselenide (CGS). These tests showed that of the three compounds, CdTe has the highest toxicity and CGS has the lowest. No comparisons were made with the parent Cd and Te substances. Cadmium, one of the CdTe precursors, is an extremely hazardous material. The acute health effects from inhaling it include pneumonitis, pulmonary edema, and death. However, CdTe is insoluble to water; as such, it may be less toxic than cadmium. This issue needs further investigation.

In solar cell manufacturing facilities, workers may be exposed to Cd compounds through inhalation of contaminated air and by ingestion from hand-to-mouth contact. Inhalation is probably the most important pathway because of the greater potential for exposure and the higher efficiency of absorption of Cd compounds through the lung than through the gastrointestinal tract. The physical state in which the Cd compound is used and/or released to the environment is another determinant of risk. Processes in which Cd compounds are used or produced in the form of fine fumes or respirable particulates present larger health risks than when they are in liquid or solid forms. Similarly, operations involving volatile or soluble Cd compounds (e.g., CdCl_2) must be more closely scrutinized. Hazards to workers may arise from preparing feedstock, fume/vapor leaks, etching of excess materials from panels, maintenance

operations (e.g., scraping and cleaning), and handling of waste. Caution must be exercised when working with this material, and several layers of control must be implemented to prevent employees from being exposed. In general, the hierarchy includes engineering controls, personal protective equipment, and work practices. Area and personal monitoring would provide information on the type and extent of workers' exposure, assist in identifying potential sources of exposure, and provide data on the effectiveness of the controls. The U.S. industry is vigilant in preventing health risks and has established proactive programs in industrial hygiene and environmental control. Workers' exposure to Cd in PV manufacturing facilities is controlled by rigorous industrial hygiene practices and is continuously monitored by medical tests, thus minimizing potential health risks.

4.2 Public Health and Environmental Issues

No public health issues have been identified for this technology. Environmental issues are related to the disposal of manufacturing waste and end-of-life modules; these are discussed in Section 8.

5. COPPER INDIUM DISSELENIDE SOLAR CELLS

5.1 Occupational Health and Safety

The most common processes for forming CIS solar cells are the coevaporation of Cu, In, and Se and the selenization of Cu and In layers in an H_2Se atmosphere. The toxicity of Cu, In, and Se is considered mild. There is little information on the toxicity of CIS. Laboratory animal studies have shown CIS to have mild to moderate toxicity in the respiratory track; in comparing CIS, CGS, and CdTe, CIS was found to be less toxic than CdTe and somewhat more toxic than CGS.

The selenium threshold level [threshold limit value, time-weighted average (TLV-TWA)] of 0.2 mg/m^3 was set to prevent systemic toxicity and to minimize the potential for irritation of the eyes and upper respiratory tract. Interestingly, selenium is an essential element in the human diet and in small doses has exhibited cancer-preventive properties.

Although elemental selenium is only mildly toxic, hydrogen selenide is highly toxic. It has an immediately dangerous to life and health (IDLH) concentra-

tion of only 1 part per million (ppm). Hydrogen selenide resembles arsine physiologically; however, its vapor pressure is lower than that of arsine and it is oxidized to the less toxic selenium on the mucous membranes of the respiratory track. Hydrogen selenide has a TLV-TWA of 0.05 ppm to prevent irritation and the onset of chronic hydrogen selenide-related disease. To mitigate hazards from H_2Se , the CIS deposition system should be enclosed under negative pressure and be exhausted through an emergency control scrubber. The same applies to the gas cabinets containing H_2Se cylinders in use.

The options for substitution, isolation, work practices, and personnel monitoring discussed for CdTe are also applicable to CIS manufacturing. In addition, the presence of hydrogen selenide in some CIS fabrication processes requires pertinent engineering and administrative controls to safeguard workers and the public against exposure to this highly toxic gas.

5.2 Public Health and Environmental Issues

Potential public health issues are related to using hydrogen selenide in facilities that employ it as a major feedstock material. The hazards can be minimized by using safer alternatives, limiting inventories, using flow-restricting valves, and employing other safety options. Emissions of hydrogen selenide from process tools are controlled with either wet or dry scrubbing. Also, some facilities have scrubbers that can control accidental releases of this gas. Environmental issues are related to the disposal of manufacturing waste and end-of-life modules; these are discussed in Section 8.

6. GALLIUM ARSENIDE HIGH-EFFICIENCY SOLAR CELLS

6.1 Occupational Health and Safety

Metal organic chemical vapor deposition (MOCVD) is the most common process for fabricating III/V high-efficiency PV cells. It employs as feedstock arsine and phosphine, which are highly toxic hydride gases. Similar to the handling of silane and hydrogen selenide, the safe use of these hydrides requires several layers of engineering and administrative controls to safeguard workers and the public against accidental exposure. Such requirements pose financial demands and risks that could hinder scaling up of the

technology to multi-megawatt levels. Part of the problem is that use of the hydrides in MOCVD is highly ineffective. Only approximately 2–10% are deposited on the PV panels because a 10–50 times excess of V to III (As to Ga) is required. Metal organic compounds are used more effectively, with their material utilization ranging from 20 to 50%. In scaling up to 10 MW/year production using MOCVD, the current designs of flat-plate III–V modules require several tons of arsine and hydrogen. These quantities can be effectively delivered only with tube trailers, each carrying 2 or 3 tons of gases. The potential consequences of a worst-case failure in one of them could be catastrophic. However, it is more likely that terrestrial systems will be concentrators, not flat plates, because the former would be less expensive to manufacture. PV cells concentrate light between 100 and 2000 times. Therefore, a possible practical strength is 500 times concentrators; the material requirements for producing such concentrators are 600 times less than those needed for flat plates, and worst consequences are equally reduced.

The best way to minimize both the risks associated with certain chemicals and the costs of managing risk is to assess all alternatives during the first steps of developing the technology and designing the facility. Studies by Komeno and coworkers have shown that these hydrides may be replaced in the future by using tertiary butyl arsine (TBAs) and tertiary butyl phosphine (TBP). There appear to be no intrinsic technical barriers to growing PV-quality GaAs with TBAs and GaAsP or GaInP₂ with TBP. However, until substitutes are tested and implemented, it is prudent to use arsine and phosphine from commercial, reduced-pressure containers. Research efforts are under way in Europe to replace hydrogen by inert nitrogen. Apparently, there is no inherent reason to prohibit such a substitution. However, since molecular hydrogen decomposes to some extent, and atoms participate in the gas-phase chemistry, the PV research community is challenged with learning how to optimize III–V growth conditions with nitrogen.

7. OPERATION OF PV MODULES

The operation of PV systems does not produce any emissions. Although some toxic materials are imbedded in PV modules, these are used in extremely small quantities (e.g., 5–10 g/m²) and they are safely encapsulated within sheets of glass or plastic. Toxic compounds cannot cause any adverse health effects

unless they enter the human body in harmful doses. The only pathways by which one might be exposed to any materials from a PV module are by accidentally ingesting flakes or dust particles or inhaling dust and fumes. The PV material layers are stable and solid, and they are encapsulated between thick layers of glass or plastic. Unless the module is finely ground, dust particles cannot be generated. All the PV materials discussed herein have a zero vapor pressure at ambient conditions. Therefore, it is impossible for any vapors or dust to be generated during normal use of PV modules.

The potential exists for exposure to toxic vapors via inhalation if the modules are consumed in a residential fire and people inhale the smoke from the fire. However, common residential fires in the United States are not likely to vaporize CdTe and GaAs layers; flame temperatures in roof fires are in the 800–900°C range. The melting point of CdTe is 1041°C, and evaporation occurs at 1050°C in open air and at approximately 900°C under nonoxidizing conditions. The melting point of CdS is 1750°C and that of GaAs is 1238°C. CIS starts evaporating at 600°C, but only a 20% weight loss was measured at 1000°C. Experiments at Brookhaven National Laboratory showed that CdTe modules with both back and front glass effectively encapsulate Cd when heated up to 1100°C.

8. DECOMMISSIONING PV MODULES

Photovoltaic modules may contain small amounts of regulated materials, which vary from one type to another. Environmental regulations can determine the cost and complexity of dealing with end-of-life PV modules. If they are classified as hazardous according to federal or state criteria, then special requirements for material handling, disposal, record keeping, and reporting increase the cost of decommissioning the modules. In doing so, the principal concern regards the presence of Cd in CdTe and CdS solar films and of Pb in α -Si modules if they contain Pb-based solder. Should these modules end in a municipal waste incinerator (MWI), the heavy metals could gasify and a fraction of the gases will be released into the atmosphere. If the MWI is equipped with an electrostatic precipitator (ESP), this fraction can be as small as 0.5%, with the balance of the heavy metals remaining in the ash. The ash must be disposed of in controlled landfills.

Should the modules end up in municipal landfills, there is the potential for the heavy metals to leach

out into the soil. The leachability of metals in landfills is characterized by two elution tests: the U.S. Environmental Protection Agency Toxicity Characterization Leachate Profile (TCLP) and the German DEV S4 (Deutsches Einheitsverfahren). Both tests assume a worst-case scenario for such potential leaching. In these tests, small pieces ($<1\text{ cm}^2$) of broken modules are suspended and rotated in an eluent for 24 h. The metals present in the eluent are then measured and compared with limits prescribed by each testing protocol. If the metals' concentration exceeds the limits, the modules are demonstrating the metals' leachability and may need to be recycled or disposed of in a hazardous waste landfill. If the metals are not leaching in excessive quantities, the modules can be disposed of in a commercial landfill. Early CdTe modules failed the TCLP and the DEV tests, while Apollo modules produced by BP Solar have been reported to pass the TCLP. In exploratory tests with a few commercial α -Si modules, some failed the TCLP limit for Pb by approximately 30%. No data were found on TCLP testing of currently produced α -Si modules. CIS modules passed the test for Se in a first set of experiments. No data were found for GaAs modules. The α -Si modules contain very little hazardous material and easily pass the test. The TCLP test is a conservative one that requires breaking up the whole module into very small pieces, whereas under normal conditions the PV layer is sandwiched between two layers of glass and reasonably isolated from the environment.

The ultimate solution to PV waste and end-of-life management is recycling the useful materials. Studies at Brookhaven National Laboratory and Solar Cells, Incorporated showed that recycling is technologically and economically feasible based on the current collection/recycling infrastructure and on emerging recycling technologies. Metals from used solar panels in large centralized applications can be reclaimed in metal smelting/refining facilities that use the glass as a fluxing agent and recover most of the metals by incorporating them in their product streams. In dispersed operations, small quantities and the high costs of transportation make this option relatively expensive. Research supported by the U.S. Department of Energy Small Business Research Initiative (SBRI) program aims the development of technologies for hydrometallurgical separation that may be used for both small-scale (in-house) and large-scale recycling. The PV industry is investigating these options as part of its proactive long-term environmental strategy to preserve the environmental friendliness of solar cells. Although the PV industry

will not face this problem on a large scale before 2020, today's choices of materials and module designs may very well set a precedent for the future. The PV industry may choose to recycle spent modules, even if there are no regulatory issues related to their disposal, because recycling may improve market penetration.

9. ENVIRONMENTAL BENEFITS OF PHOTOVOLTAICS OPERATION

Carbon dioxide emissions are zero during the operation of PV systems because they require little or no maintenance or oversight. Some CO_2 emissions can be attributed to manufacturing because it takes energy to make a module. Thus, PV systems require some input energy, which they pay back early in their lifetime. This energy payback (EPB) was the focus of studies by Alsema, Frankl, Kato, and others. In the past, PV manufacture had a high EPB because the technology was immature and burdened by energy-intensive processing. Existing PV systems have an EPB of 3–10 years; future technological options likely will reduce the system's EPB to 1–3 years. In this range, the amount of CO_2 displaced by PV during a 30-year lifetime outdoors is 90–97% compared to the CO_2 of the energy it offsets. For example, assuming that U.S. energy generation causes 160 g carbon equivalent of CO_2 per kilowatt-hour of electricity, then a 1-kW PV array in an average U.S. location would produce approximately 1600 kWh each year and 48,000 kWh in 30 years. This array would avoid approximately $(48,000 \times 0.95) \times 160\text{ g/kWh}$, or approximately 7 metric tons of carbon equivalent during its useful life. Building integrated photovoltaics may further shorten the period of EPB by avoiding the energy required for structural supports. Exact predictions require a knowledge of the mix of energies with which PV will be manufactured and the mix that it will displace (which will vary with location, application, and date). The U.S. Photovoltaic Industry Roadmap projected that by 2020 the domestic PV industry could provide up to 3.2 GW of new peak electricity, increasing to approximately 25 GW by 2030. These projections correspond to an average growth in PV capacity of 25% per year. The estimated mixture of applications is one-half AC distributed generation, one-sixth AC wholesale grid generation, and one-third DC and AC value applications.

Fthenakis and Morris used the MARKAL MACRO model to test these projections. MARKAL is a multiperiod, linear programming model that

captures the complex interrelationships of energy systems and their environmental emissions across the spectrum from energy resources to energy services. The model minimizes the total cost of the energy system over a multiyear period, subject to constraints, such as limitations on pollution emissions or on the growth rates of technologies. The analysis includes the entire energy system: extracting the resources; processing operations (such as refineries); electric, gas, and oil transmission and distribution operations; and end-use technologies, such as lighting, industrial heat production, and transportation technologies (from automobiles to ships and airplanes).

All energy resources on both the supply and demand sides, as well as energy efficiency and conservation technologies, compete in an even-handed manner. The model and its associated database are continually upgraded. Current projections are based on the 2001 MARKAL database, designed to closely match the U.S. Energy Information Administration's "Annual Energy Outlook 2001" except for the PV and wind data, which were obtained from the Department of Energy (DOE) and the Electric Power Research Institute (EPRI). There are 922 technologies in the MARKAL database, including demand devices such as residential heating, commercial refrigeration, and lighting. The database includes 52 electricity-generating technologies. MARKAL also projects pollution emissions based on the characteristics of the individual technologies in the energy mixture.

Table III shows the PV input data for the model. This information derives from the PV Roadmap for 2000–2020 and from the DOE/EPRI for 2025 and 2030 (the Roadmap stops at 2020). We assumed that the peak electricity demand largely overlapped the peak power demand. The only constraint on the model was the upper bound on growth of 25% per year. An average U.S. insolation of 1800 kW/h/m² was assumed; the PV penetration in regions of high solar insolation could be much greater than the average (Table IV).

Carbon and other gas emissions are an important part of the model's output. PV technology under the predicted rates of implementation in the United States would accelerate a steadily increasing displacement of carbon emissions. In 2030, this displacement is projected to be 17 million tons/year, and during the entire 30-year period PV would displace a total of 128 million metric tons of carbon (Table V).

In addition, in 2030 alone, PV would displace 169,000 tons of SO₂, 103,000 ton/year of NO_x, and 6200 tons of particulates (Table VI).

TABLE III

Cost Assumptions for PV (\$/Wp) Input Values in MARKAL

	2000	2005	2010	2015	2020	2025	2030
\$/Wp	6.1	4.3	3.0	2.1	1.5	1.0	0.9

TABLE IV

Total U.S. Capacity and % PV for the MARKAL Roadmap Case

	2000	2005	2010	2015	2020	2025	2030
Total capacity (kW/h/m ²)	768	821	895	972	1033	1129	1252
PV capacity (kW/h/m ²)	0.08	0.32	1.19	3.84	11.93	36.6	111.9
% PV	0.01	0.04	0.13	0.40	1.15	3.24	8.94

TABLE V

Displacement of Carbon Emissions by PV (Million Metric Tons/Year)

	2000	2005	2010	2015	2020	2025	2030
Carbon displaced	0.02	0.06	0.21	0.64	1.98	5.80	17.02

TABLE VI

Displacement of Other Emissions by PV (Thousands of Tons/Year)

	2000	2005	2010	2015	2020	2025	2030
NO _x	0.17	0.70	2.23	5.92	15.60	39.8	103
PM10	0.005	0.02	0.08	0.22	0.89	2.10	6.22
SO ₂	0.23	0.82	2.62	7.45	21.3	59.4	169

This analysis addresses only the grid-connected portion of the Roadmap in the highly competitive U.S. market; it does not examine the PV prospects of the promising markets of the developing world.

10. CONCLUSION

The manufacture of PV modules uses some hazardous materials, which can present health and safety risks if adequate precautions are not taken. Routine conditions in manufacturing facilities should not pose any threat to health and the environment.

Hazardous materials could adversely affect occupational health and, in some instances, public health during accidents. Such hazards arise primarily from the toxicity and explosiveness of specific gases. Accidental releases of hazardous gases and vapors can be prevented by using safer technologies, processes, and materials, better material utilization, and enhancing employee training and safety procedures. As the PV industry vigilantly and systematically approaches these issues and mitigation strategies, the risk to the industry, the workers, and the public is continually minimized.

Photovoltaics in the energy market will replace fossil fuel combustion technologies and will subsequently reduce the discharges of carbon dioxide, sulfur dioxide, nitrogen oxide, and particulates throughout the region where they are installed. Based on energy–environmental–economic analysis, we predict that with the current price projections for the energy technologies in the U.S. energy mixture, photovoltaics could reach a capacity of 12 GW by 2020 and 112 GW by 2030, eliminating a significant amount of pollutants from the U.S. energy cycle.

SEE ALSO THE FOLLOWING ARTICLES

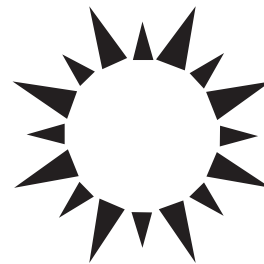
Arid Environments, Impacts of Energy Development in • Electricity, Environmental Impacts of • Photovoltaic Conversion: Space Applications • Photovoltaic Energy: Stand-Alone and Grid-Connected

Systems • Photovoltaic Materials, Physics of • Solar Cells

Further Reading

- Alsema, E. A. (1998). CO₂ and energy payback of PV. In “Photovoltaics and the Environment” (V. Fthenakis, *et al.*, Eds.), Workshop Report BNL-52557. Brookhaven National Laboratory, Upton, NY.
- Frankl, P., and Gamberale, M. (1998). Analysis of energy and CO₂ aspects of building integration of photovoltaic systems. In “Photovoltaics and the Environment (V. Fthenakis, *et al.*, Eds.), Workshop Report BNL-52557. Brookhaven National Laboratory, Upton, NY.
- Fthenakis, V. M. (2000). End-of-life management and recycling of PV modules. *Energy Policy* 28, 1051–1058.
- Fthenakis, V. M. (2001). Multi-layer protection analysis for photovoltaic manufacturing. *Process Safety Prog.* 20(2), 1–8.
- Fthenakis, V. M., and Gonsiorawski, R. (1999). Lead-free solder technology from ASE Americas, Workshop Report BNL-67536. Brookhaven National Laboratory, Upton, NY.
- Fthenakis, V. M., and Morris, S. (2003). MARKAL analysis of PV capacity and CO₂ emissions’ reduction in the U.S. In “Proceedings of the World Conference on Photovoltaic Energy Conversion (WCPEC-3), May 11–18, 2003, Osaka, Japan.”
- Fthenakis, V., Morris, S., Moskowitz, P., and Morgan, D. (1999). Toxicity of cadmium telluride, copper indium diselenide, and copper gallium diselenide. *Prog. Photovoltaics* 7, 489–497.
- Kato, K., Hibino, T., Komoto, K., Ihara, S., Yamamoto, S., and Fujihara, H. (2001). A life cycle analysis on thin film CdS/CdTe PV modules. *Solar Energy Materials Solar Cells* 67, 279–287.
- Komeno, J. (1994). Metalorganic vapor phase epitaxy using organic group V precursors. *J. Crystal Growth* 145(1–4), 468–472.
- Tamanini, F., Chaffee, J. L., and Jambar, R. L. (1998). Reactivity and ignition characteristics of silane/air mixtures. *Process Safety Prog.* 17(4), 243–258.

Physics and Economics of Energy, Conceptual Overview



KARL-ERIK ERIKSSON

Karlstad University
Karlstad, Sweden

-
1. Introduction
 2. Physical Aspects of Energy
 3. Cosmic Origin of Exergy Resources
 4. Energy in Society
 5. Energy, Values, and Economics
 6. The Evolutionary Perspective

Glossary

entropy A measure of the disorder or unavailability of energy within a closed system. More entropy means less energy is available for doing work.

First Law of Thermodynamics The internal energy in a system isolated from other systems remains constant (energy is conserved).

Law of Conservation of Matter During an ordinary chemical change, there is no detectable increase or decrease in the quantity of matter.

Second Law of Thermodynamics The principle that the disorder of an isolated system (entropy) increases to a maximum.

1. INTRODUCTION

The use of energy in human society is a many-faceted subject. Many fields of knowledge are involved in the understanding of this subject. Energy is a fundamental concept in physics. When dealing with energy in society, a certain usefulness is assumed, a set of qualities of energy forms that are also possible to describe in physical terms. Energy is procured from natural systems and given back to natural systems after transformations in the society. These processes are economic and biological as well as technical in nature and they are often regulated by society. The technical processes can be designed or analyzed on

the basis of the physical sciences. Exchanges and regulations can be analyzed in terms of values and costs/risks. Here, environmental and health aspects become intertwined with technical, economic, and judicial aspects. On a fundamental level, there are the value aspects, in particular, ethical aspects.

One must keep in mind that all these aspects are relevant. In science, one often must simplify, and there are often temptations for scientists to adopt one aspect as being supreme and to let it dominate the description or analysis. Simplified pictures of this kind can be very pedagogical, but they may also lead to misconceptions that can be very harmful. One way to check one's understanding is to ask what the system/process or the problem in question looks like on another level. Traditionally, in science, analyzing in terms of subsystems or subprocesses has been a very successful approach. It has become increasingly clear that it may also be necessary to analyze a certain system or process in relation to the larger context, sometimes several larger contexts, in which it is embedded.

In the case of energy in society, from looking mainly at technical systems for providing primary energy, one has been forced to ask whether the whole societal energy system can be sustainably embedded in the ecosphere. For nuclear energy, one has been forced to ask whether the bulk handling of radioactive material can be performed safely in the present and future world with its conflicting interests and unforeseeable risks. The time scale involved in considering the handling of nuclear waste is long not only compared to lead times in energy technology but even compared to the known human history.

One system boundary that must be broken immediately is the one between energy and matter. First, most of the crucial problems of the energy system are material-related. In the case of energy and climate, it is the transition of carbon from its original

state in fossil coal, oil, or gas to atmospheric carbon dioxide that is crucial. Second, a very large part of the energy use in society is related to material transformations or transports of materials. Energy use in the aluminum and cement industries are two examples of this.

Physical processes of conversions of energy and materials are fundamental for the metabolism of living organisms and for the metabolism of human society. Different forms of energy and matter have different physical (thermodynamic) values and also different degrees of usefulness and economic value.

Sections 2 and 3 deal with basic concepts and principles related to the physical aspects of energy. Section 4 deals with the use of energy in society and the interface between physics and economics. Section 5 discusses energy in connection to values and economic principles. The two aspects are brought together in Section 6, where the evolutionary perspective on the use of energy and matter meets the human responsibility of today's resource handling on Earth.

2. PHYSICAL ASPECTS OF ENERGY

The scientific field that deals with transformations of energy and materials is thermodynamics, which describes macroscopic quantities of energy and matter. The theoretical way to understand this is statistical; the underlying theory is statistical mechanics, which analyzes populations of many particles and their motions and interactions.

In the late 1950s, E. T. Jaynes showed how statistical mechanics can be interpreted in terms of another relatively new theory, C. E. Shannon's theory of information, basically a theory of signals. It deals with probability distributions, conjectured or observed. Information theory is of fundamental interest for describing an important use of energy in nature and society, the build up of structures.

In information theory, an answer, "yes" or "no," or, equivalently, a binary figure, 0 or 1, carries one bit of information, against a background of total uncertainty with *a priori* probability $\frac{1}{2}$ for 0 or 1. The information carried by a certain symbol having a priori probability p is (in bits)

$$^2\log\frac{1}{p}.$$

Clearly, when $p = \frac{1}{2}$, the result is unity.

In a text, the different letters appear with different probabilities (different frequencies), and hence they

carry different amounts of information. In an n -letter alphabet, the letters can be denoted by i ($i = 1, 2, \dots, n$). Let p_i be the probability that, in a randomly chosen space, the letter that one finds there is i . Weighing the information gain in finding the letter i , $^2\log\frac{1}{p_i}$ with the probability p_i for its appearance, and summing over all letters, one obtains the expected information gain, i.e., the present information deficit,

$$S(p) = \sum_{i=1}^n p_i ^2\log\frac{1}{p_i}.$$

This is the entropy of the probability distribution $p = \{p_i\}$.

From looking just at single letters in a text, one can continue and look at pairs of neighboring letters or, more generally, at strings, i.e., sequences, of increasing length. In a text in a certain language, there are correlations between neighboring letters; i.e., certain letter combinations occur more frequently than others.

If there are two probability assignments, $p^{(0)} = \{p_i^{(0)}\}$ and $p = \{p_i\}$, and the latter is based on more knowledge than the former, then the information gain in going from the former to the latter is

$$\begin{aligned} K(p^{(0)}; p) &= \sum_{i=1}^n p_i \left[- \left(^2\log\frac{1}{p_i} - ^2\log\frac{1}{p_i^{(0)}} \right) \right] \\ &= \sum_{i=1}^n p_i ^2\log\frac{p_i}{p_i^{(0)}}. \end{aligned}$$

(The information difference is minus the information deficit difference; moreover, the best probability assignment $p = \{p_i\}$ must be used to give the weights.)

Here, $K(p^{(0)}; p)$ is the relative entropy (or Kullback information) of the probability distribution p with respect to the reference distribution $p^{(0)} = \{p_i^{(0)}\}$. The Kullback information has the property that it is never negative and it is zero only when the probability distributions coincide.

These two concepts, entropy and relative entropy, are the basic concepts of information theory. They are very flexible and can be readily used in information theory, statistics, linguistics, and physics, as well as in other sciences.

As already mentioned, one may generalize from a string of discrete letters, i.e., a text, to a lattice in two or more dimensions. An important example of a two-dimensional lattice is a picture image, consisting of pixels, in black/white or color. Generalizing to continuous systems in any number of dimensions is also possible. In classical physics, one has vectors in the so-called phase space of positions and momenta of many particles, and the resulting theory is, as

mentioned already, statistical mechanics. Here, probability assignments represent macroscopic physical states. Extension into quantum mechanics is quite straightforward but is not discussed here. It should be mentioned, though, that this extension is essential for several important energy technologies.

Entropy is thus a fundamental quantity not only in information theory but also in statistical mechanics and thermodynamics. The two entropies are identical for a physical system except for a constant factor,

$$S_{\text{thermodyn}} = (k \ln 2) S_{\text{info}},$$

where k is Boltzmann's constant, and

$$k \ln 2 = 1.0 \times 10^{-23} \text{ JK}^{-1} \text{ bit}^{-1}.$$

Energy is a fundamental physical quantity. It is closely related to time and momentum is closely related to space. In fact, the quantity representing energy in quantum mechanics also has the function of generating displacement in time and momentum generates spatial displacement. Conservation of energy, the first law of thermodynamics, is related to the fact that the laws of nature are invariant under time displacements. Similarly, conservation of momentum (the law of inertia) and conservation of angular momentum follow from the universality of physical laws in space.

Matter, under the conditions of the universe of today, is made up of nucleons (neutrons and protons, the constituents of atomic nuclei) and electrons. Three quarks constitute a nucleon. The mass of an electron is approximately 1/200 of the mass of a nucleon. In particle physics, as studied at accelerator laboratories, a large variety of particles made up of quarks and anti-quarks are created and investigated. This includes the formation of matter and antimatter. Also leptons, the particle family to which the electron belongs, can be similarly produced and studied.

Electric charge is conserved. The same is true for lepton number and quark number with extremely good accuracy (except possibly for extreme situations, such as in the very early universe). Since accelerator laboratory reactions are rare, this means that for all practical purposes, one can regard the number of nucleons as constant and consider this the law of conservation of matter.

The state of a system of many particles is usually not exactly known. Even if the laws of motion were exactly known, the uncertainty would still grow in time due to delicate dependence on unknown parameters of the initial state. Hence, for any system

left by itself, the entropy increases. This is the second law of thermodynamics.

When the entropy of a closed system has reached its maximum value, the system will remain in this state; i.e., it is in its equilibrium state. Hence, the deviation of entropy from its maximum value is a measure of the degree of nonequilibrium.

Another useful measure related to this is the Kullback information or relative entropy, defined above. Where there is disequilibrium, there is, at least in principle, the possibility of extracting work. Consider an ideal process, i.e., a reversible and entropy non-producing process, to extract mechanical work (which is of zero entropy) from a system, while it approaches equilibrium. Then, the amount of extractable work from the system, called the exergy E of the system, is proportional to the Kullback information,

$$E = (k \ln 2) T_0 K(p^{(0)}; p).$$

Here, p is the original state of the system, and $p^{(0)}$ is the equilibrium state of temperature, T_0 , reached reversibly under the extraction of the amount E of mechanical work.

This may all seem very theoretical, but it is applicable to all sorts of disequilibria, such as differences in temperature or in pressure or chemical disequilibria. It is thus generally applicable to physical and technical systems and processes for energy and material transfers and conversions.

Earth is situated in a radiation flow from the Sun and it radiates this energy as heat into the interstellar space. The geophysical systems and life on Earth's surface convert radiated energy coming from the Sun at an average temperature of 5800 K. Approximately 30% is reflected. After conversions, involving the water cycle and the carbon cycle, the absorbed energy leaves Earth as radiation at an average temperature of 255 K (−18°C). This can be viewed as a huge heat engine with an available exergy power of 115,000 TW. Of this, the global society uses less than 0.01%.

In terms of information inflow or, rather, flow of information capacity, the radiation exchange of Earth amounts to 4×10^{37} bits per second. Of the huge inflow of information, some has been coded into the structure of life and the complex molecular interplays within the cell, also into the regulatory mechanisms within organisms, between individuals of a population, and between species of an ecosystem. Life has interacted with and changed the geophysical systems in a mutually adaptive evolution. Earth has a large system of biogeochemical

cycles on which life depends. Except for the influence of actions of the human society, in particular the industrial society, this system would be relatively stationary.

There are many examples of structure formation in systems with a through-flow of energy (Fig. 1). The whole living world contains an enormous richness of such phenomena on various scales. There are also many non-living examples, such as cloud formation and the formation of snowflakes.

A classroom experiment of self-organization, easy to repeat, is the set of reactions in a solution of chemicals, devised by Belousov and Zhabotinsky. The solution contains an energy-rich substance and substances that become involved in nonlinear reaction kinetics and diffusion. Small thermal fluctuations in concentrations are sufficient to trigger the system to form visible geometrical structures that are slowly changing and eventually decaying.

Figure 2 shows a diagram of exergy (information capacity) flows in such a self-organizing system. The same principle applies also to spontaneous structuring on various levels in cosmic, stellar, and biological evolution. A through-flow of energy can give a net inflow of exergy. In a self-organizing system, there are flows driven by gradients, differences in concentration, pressure, or temperature. Such motions are irreversible processes involving entropy production.

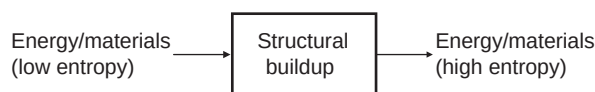


FIGURE 1 Diagram of structure formation of a system in a flow of energy or of energy with matter. Due to differences in entropy between the flows, there is a net inflow of exergy or information capacity. There is also entropy production (dissipation) within the system.

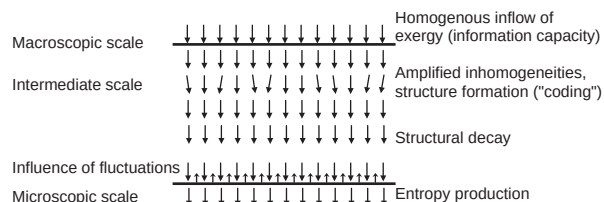


FIGURE 2 Diagram indicating flow of information from the macroscopic scale to the microscopic scale via an intermediate scale with structure formation triggered by fluctuations. Structure formation can take place on different scales. Structure formation and decay are both present in the flow from macro to micro.

In practice, of course, no process is totally ideal/reversible; there is always production of entropy. Therefore, not all of the exergy present in a system can be extracted as work. The exergy loss in a conversion process is proportional to the entropy produced in the process,

$$W = E - T_0 \Delta S.$$

(Here, E the original exergy, W is the extracted work, ΔS is the entropy production, and T_0 is a reference temperature, usually the ambient temperature or the equilibrium temperature.)

Thus, exergy can be lost but not created. It is something valuable, also economically. Therefore, it is useful as a physical resource measure. Expressed in thermodynamic quantities, the exergy is

$$E = U - T_0 S + p_0 V - \sum_i \mu_{i0} N_i,$$

where U is the internal energy, S is the entropy, V is the volume, and N_i are the amounts of different chemical components. Furthermore, T_0 is the reference temperature, p_0 is the reference pressure, and μ_{i0} are reference chemical potentials. The exergy is thus context-dependent through the choice of reference system.

Apart from nuclear reactions and decays, chemical elements are conserved. Thus, chemical elements, energy, and exergy can be subject to bookkeeping; i.e., balance accounts can be set up for inflows, outflows, and internal movements and conversions. Clearly, such bookkeeping can easily be generalized to include nuclear reactions and decays, for instance, in the nuclear industry. If the context dependence of exergy is to be avoided, entropy and entropy production can be taken into account instead. Physical bookkeeping can be done for natural systems, such as an ecosystem, a population, an individual organism, or a cell, but also for production/consumption units, such as a factory, a machine, a car, or a household.

The disequilibrium between the dense and hot Sun and the cold interstellar space functions like the poles of a large battery. Earth is situated in a radiation flow, an energy current from the hot pole to the cold pole of this battery. On Earth, this current charges secondary batteries and secondary currents: temperature and pressure differences, differences in gravitational potentials between high-altitude dams and the sea; the water cycle, including rivers and ocean currents, winds, and waves. Many of those secondary batteries and currents are used by the societal energy systems.

3. COSMIC ORIGIN OF EXERGY RESOURCES

It may be natural to ask for the origin of the Sun/space battery. How was it charged? The answer is that two disequilibria developed in the early universe, one nuclear and one gravitational. Both were related to the expansion of the universe and with this the changing equilibrium conditions (and with this the maximum possible entropy).

According to the standard theory of the early universe, nuclear matter existed as nucleons and was in thermal equilibrium 0.01 s after the Big Bang. The universe continued to expand and when the temperature passed below the dissociation temperatures for ordinary nuclei, the possible equilibrium state for nuclear matter became iron nuclei (^{56}Fe). At that time, approximately $\frac{1}{8}$ of the nucleons were neutrons; the rest were protons. Only the first steps of nuclear reactions could take place under the prevailing conditions. During the fourth minute, neutrons joined with protons to form helium (^4He). After this, almost all nuclear matter was in the form of helium ($\frac{1}{4}$) and the remaining hydrogen, i.e., protons ($\frac{3}{4}$). The conditions needed for continued nuclear reactions were not prevalent. The nuclear fuel of the early universe was thus saved for later burning in stars. During the first hours of the universe, through the cooling by cosmic expansion, the difference widened between the actual state of nuclear matter and the equilibrium state. The nuclear matter had become “supercooled” through cosmic expansion and nuclear exergy had been created.

In the early universe, matter was suspended in radiation. When the temperature dropped below the ionization temperature of helium and hydrogen a few hundred thousand years after the Big Bang, neutral atoms could form and radiation became more weakly coupled to matter. Then, gravitational clumping was possible and gravitational exergy was created. Matter could assemble on various levels to become galaxies or solar systems. Stars could start to form and in them, energy was converted into radiation through nuclear burning, which produced the various chemical elements. Matter thrown out from exploding stars was reassembled in later generations of stars, such as the Sun. What now remains of the original radiation is the 3 K cosmic background radiation.

With self-organizing systems, the old clockwork metaphor of physical science had to give way to a new view of dynamics, supported by a vastly increased capacity to study processes through

computer simulations. The clockwork itself is a dissipative (entropy-producing) system: when it runs down, it converts mechanical energy into heat. Even if fundamental physical laws, as in quantum field theory, may be elegantly summarized in a principle of least action, actual physics is still indeterministic due to uncertainties in initial conditions and external influences. The old deterministic way of viewing physics may have had too strong an impact on other sciences, providing them with obsolete ideals.

4. ENERGY IN SOCIETY

It is of interest to consider the main uses of exergy in society. They can be summarized as follows:

1. to replace losses that occur through
 - i. leakages (e.g., heat leaking in or out through roofs or walls),
 - ii. friction (e.g., in machines or vehicles), and
 - iii. material degradation and dispersion (through wearing and tearing and corrosion); and
2. to steer irreversible processes of flows from one state to another (e.g., matter moving from high to low pressure or thermal energy moving from hot to cold).

With this kind of description, the places to look for improved efficiency become more obvious.

Physical bookkeeping, as described above, is then of interest for analyzing the efficiency of a system or for studying a system as part of a larger system, such as a community as part of an ecosystem in which it is embedded. In the 1970s, Robert Ayres developed physics-based principles and schemes for book keeping of the turnover of energy and materials in society.

Human society is a major component in Earth's cycles of nitrogen and sulfur. Humans are systematically interfering with the carbon/oxygen flows by bringing fossil carbon into the atmosphere as CO_2 . Thereby, the natural greenhouse effect that makes Earth inhabitable is enhanced, resulting in global warming and risks for major climatic changes.

Some material flows are crucial for the functioning of the ecosphere. Others are toxic or geophysically harmful and must be avoided. This means that, in material balance accounts, some flows must be more carefully guarded than others. There is no single quantity that can function as an overall index. Good overviews are needed, but the generalists cannot do away with the detailed knowledge provided by the specialists.

Robert Ayres also pointed out that energy and matter are never consumed; both remain after consumption. Even the matter of our bodies is continuously exchanged; it is a temporary loan to be ultimately paid back. Exergy, i.e., energy quality, is consumed, however, through entropy production.

What is useful is a flow of services that is enjoyed, directly or in the use of products. Thus, “product” is an intermediary concept; the concept of basic economic importance is “service.” It is useful to think of temporary leasing of products (carrying energy and matter) as a kind of service being consumed.

Figure 3 shows a diagrammatic representation of one unit in societal resource use and service production. In Fig. 4, production of a displacement service (transport or storage) is similarly represented.

Strictly speaking, a production process is not completed until all the flows have been duly processed in society or nature. This means that in the case of fossil fuels, the relaxation mechanisms of the atmosphere and climatic systems are to be viewed as part of the process of producing energy services. This is anomalous in the sense that completion of the production process must continue long after the consumption of the energy services. The waste handling in the nuclear energy industry, as mentioned, is anomalous in the same way.

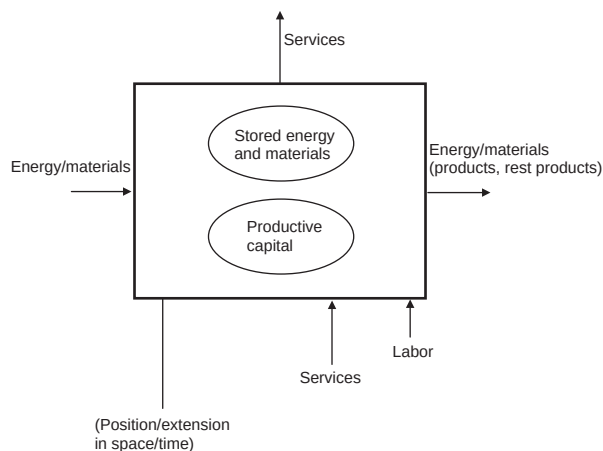


FIGURE 3 The metabolism of a society can be viewed as being built up of transformation activities like the one represented here. Physical flows are shown as horizontal arrows. Primary inflows come from natural systems and the final physical outflows also end up in natural systems. Only services are finally consumed. Products are viewed as intermediate entities. Leasing of a product over a time period is viewed as a service. Here labor is represented as nonphysical, since its human and economic aspects are more important than its physical aspects.

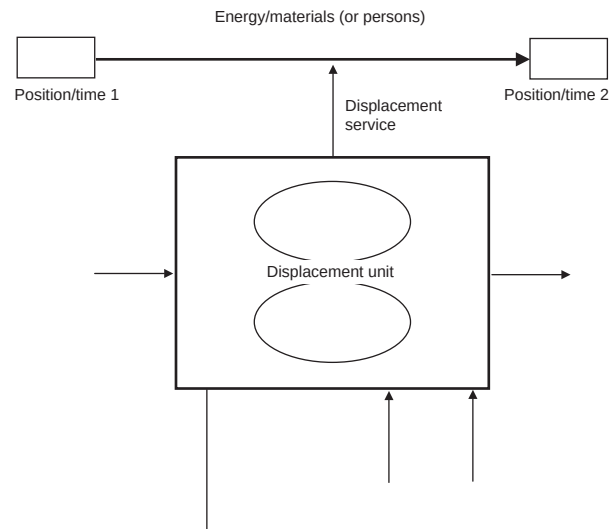


FIGURE 4 Diagram of a transport or storage (displacement) production unit, providing a transport or storage service.

Exergy was noted above as a resource measure. It is very general but still it cannot serve as a total resource measure, because it does not include all important aspects of the actual context. Exergy is needed to run everyday systems that make bodies, homes, and societies function. It is used (and most of it is consumed) for both maintenance and buildup.

Sadi Carnot’s classical problem of maximizing the work output from heat can be viewed as an engineering problem with a physical solution but also as an economization problem. Thus, exergy expresses thermodynamic value but can also carry economic value.

To determine the space available for the totality of human activities (Daly’s issue of scale, a reasonable size of one’s ecological footprint or environmental space), one must analyze how the global society as a physical and ecological system can be sustainably embedded in the ecosphere. The global society is clearly far from static. A more just distribution of resources involves expansion, which must be compensated for by structural changes and increased efficiency (“factor 4” or “factor 10”).

Very important discussions are going on in connection to the United Nations system, United Nations conferences and international negotiations, as well as in environmental science and in development research and in non-governmental organizations working with these issues. Consumers’ responses to environmental and fair-trade rating of products have taken global issues into everyday life.

As mentioned above the exergy flow into Earth through its radiation exchange (Sun/space) is vast. There are large areas where solar energy could be efficiently harvested, in areas where people live as well as in uninhabited areas. Hydrogen can be expected to become a light and convenient energy carrier. The amount of exergy used in human society could be collected by photovoltaics in a rather limited desert area. The transition from fossil fuels to solar energy is at least physically feasible.

From the 2002 United Nations Conference on Environment and Development in Johannesburg, South Africa, the Swedish Ministry of Environment summarizes the discussion within the Forum on Science, Technology, and Innovation for Sustainable Development as follows (translation from Swedish by K.-E.E.):

From frontline research it is reported that it is possible to develop the technologies needed for the whole of human-kind to enjoy a high standard of living. Access to clean water and good sanitation, clean energy, food and good health conditions for everybody is totally feasible. In this development, one must take as points of departure the ecological systems that form the basis for all this. Research and new investments will be needed. Sufficient amounts of capital and other resources are available but have to be redistributed—a massive resource mobilization is necessary from the public sphere as well as the private sphere.

—Kjellén et al., 2002, Summary

It would be good to have a clearer picture of a future sustainable resource use technology and the action space allowed for, say, an average family in a just global society.

The movement away from dumping CO₂ into the atmosphere may for some time involve sequestering of CO₂ into underground deposits, but in the long run, the necessary decarbonization must take place through abandoning of fossil carbon and a transition to renewable primary energy.

It is worth noting that, apart from biomass, renewable energy leads not only to decarbonization but also to dematerialization by largely removing material energy carriers. Wind energy and solar energy can be harvested without anything that corresponds to hydropower dams or nuclear reactors, and when a material energy carrier is needed, it can be hydrogen. More sophisticated material quality management can be achieved through repair, reuse, and recycling and through substitution of organic materials and light metals for heavier metals.

It is important to note that several materials that are candidates for energy converters and energy storage (photovoltaics, fuel cells, and batteries) are

problematic. Some are quite rare in Earth's crust; some have toxic effects and should not be allowed into the ecosphere. Fortunately, silicon, still a top candidate for photovoltaics, is both abundant and nontoxic. Exactly which technological combinations will prove sustainable in the long term cannot be foreseen as yet.

One can ask what the resource system of a just global society would look like and how it could be organized administratively and economically to serve the average family mentioned above. It should be possible to use physical, technical, and ecological constraints as inputs in economic models and deduce relative prices, probably quite different from today's prices.

5. ENERGY, VALUES, AND ECONOMICS

We have seen that exergy is basic for life and activities in nature and society. Exergy is a measure of value in physical terms. However, the more complex a system or a process becomes, the less can be said by physics. Even the term “complexity” is problematic. There are several definitions of complexity as a quantitative concept in information theory. These definitions capture important aspects, but there is no mature theory of complexity. For the understanding of complex systems and processes, one needs the life sciences, social sciences, and humanities. What about economics, in particular, energy economics?

Some years ago, in a draft of the economics chapter for a report from the Intergovernmental Panel on Climate Change (IPCC), in an estimate of the costs of climate change, a statistical life in a poor country was valued as $\frac{1}{15}$ of a statistical life in a rich country. Willingness to pay for life insurance was part of the basis for this finding.

This led to very strong criticism by many scientists and finally the discussion convinced the authors that a change was necessary in the principles for making cost estimates in a global perspective. Willingness (or rather ability) to pay for life insurance could no longer be considered a reasonable basis for value. The presently dominant ideology that attributes a fundamental significance to market and money was shaken, at least temporarily. If the IPCC draft had taken into account that the marginal value of money is greater for a poor person than for a rich person, then a factor of approximately 15 would have appeared and canceled the factor $\frac{1}{15}$.

All this shows that moral values can be deeply involved in dealing with economic evaluations of energy systems.

If the idea that economic value is tied to human welfare is taken seriously, then this has very strong implications. The global welfare deficit can be computed for the current very unequal income distribution with a conventional welfare function. The deficit can then be expressed as a cost in money terms, using money value at the average income level. The cost turns out to be on the same order of magnitude as the total global income or larger. Thus, if inequality were included in measures of economic performance, economic descriptions would change drastically.

This discussion shows that it is important to test the validity limits of reasoning and concepts. It is especially so with economics, considering the dominant position of economic rationality for decision-making on almost all levels.

An important study in this context attempted to arrive at an economic value of nature. This work showed that large modifications are needed in economics in order to include the value of natural systems. It also showed that the uncertainties are large. Economics also has difficulties in handling not only natural capital, but also human capital and social and cultural capital and the value flows associated with these.

Either such value aspects should be included in economic considerations or those value aspects should define the boundaries within which economic considerations are to be made. As the above example from the IPCC draft shows, pure economic considerations, not governed by human value, can become immoral. The conclusion is that physical and economic aspects are both very important, but one should be well aware of their limitations.

Efficiency is a topic where physics and economics meet. Carnot's efficiency problem opened a new branch of physics: thermodynamics; at the same time, it was an economic problem. Thermodynamic efficiency can be economically visible through an exergy price.

There are also other links. Distributional inefficiency is related to technical (and thermodynamic) inefficiency: poor people may not afford investments to improve efficiency and rich people may not feel the fees introduced to discourage them from waste. Organizational efficiency and social efficiency are also relevant aspects to be included in an efficiency analysis.

Justice between generations can be seen as an intertemporal efficiency. If the time anomalies in

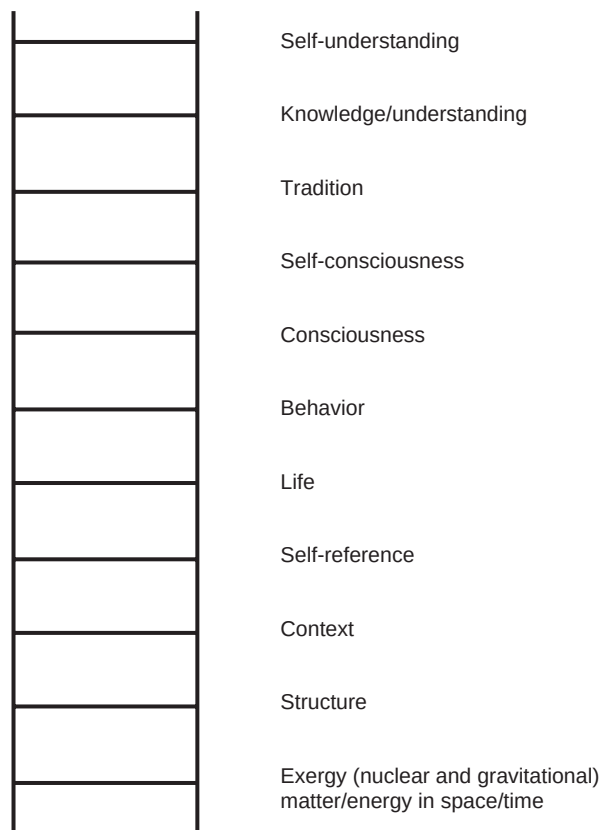


FIGURE 5 Diagram of an evolution ladder with emergent system characteristics.

production described above are to be taken seriously, then, for consistency, there should be no discounting of the future; the *a priori* discount must be zero.

6. THE EVOLUTIONARY PERSPECTIVE

Since the very early beginnings of the universe, transformations of energy and matter have involved the emergence of new systems and phenomena. The history of the universe is a history of creation. Structure was created or became manifest in early clouds on a different scale, to become solar systems or galaxies.

The surface of Earth developed conditions favorable for the richness of carbon chemistry, a wealth of possible forms and processes of varying permanence and flexibility. Self-organization also came to involve self-reference, i.e., systems carrying blueprints of themselves. Accidental changes in such blueprints sometimes led to more efficient modified systems. Reproduction, life, variation, selection, and

evolution became possible. Evolution is a process of accumulating information, selected on the basis of its functional advantages for reproduction. The story is as represented in Figs. 1 and 2, repeated on many levels and many scales. Information capacity (exergy) became coded into relatively stable information carrying a meaning.

The ladder in Fig. 5 is an attempt to identify steps in evolution, starting with the cosmic origin of the universe in the Big Bang. The steps may be quite ambiguous, but they are meant to convey the evolutionary perspective of modern science. We human beings may be seen as organs, through which the universe/cosmos can view itself and reflect over itself.

Not knowing of evolution in other parts of the universe, it is reasonable to ask what evolutionary step could or should be taken by humans on Earth. Information and communication technology reaches almost everywhere on Earth and is becoming global in character. Technically, humankind is one unit. Ecologically, we humans are strongly dependent on one another in the treatment of Earth. However, economically, politically, and ideologically, we are still divided. In the quoted summary from Johannesburg, adequate resources for a good life for everybody, including clean energy, is described as a realistic possibility. The next evolutionary step on the ladder would therefore be to unite, to develop and share a common self-understanding, and to act together and organize a good life on Earth.

SEE ALSO THE FOLLOWING ARTICLES

Complex Systems and Energy • Earth's Energy Balance • Economic Thought, History of Energy in

• Entropy and the Economic Process • Evolutionary Economics and Energy • Exergoeconomics • Exergy Analysis of Energy Systems • Information Theory and Energy • Input–Output Analysis • Thermodynamics and Economics, Overview • Thermodynamics, Laws of • Work, Power, and Energy

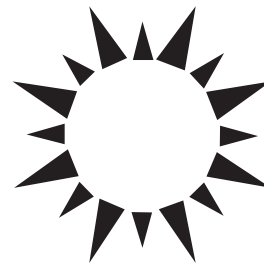
Further Reading

- Andersson, B. A., and Jacobsson, S. (2000). *Energy Policy* 28, 1037.
- Daly, H. E., Cobb, J. B., Jr. (1989). "For the Common Good." Beacon, Boston, MA.
- Eriksson, K. E. (1994). *Ric. Econ.* 48, 341
- Eriksson, K. E. (2002). Lecture notes from the 4th Regional College on Modelling, Simulation, and Optimization, Accra, Ghana, 2002, National Centre for Mathematics and Science, Accra, Ghana.
- Eriksson, K. E., Islam, S., and Skagerstam, B. S. (1982). *Nature* 296, 540.
- Eriksson, K. E., Lindgren, K., and Månsson, B. (1987). "Structure, Context, Complexity, Organization." World Scientific, Singapore.
- Holmberg, J., Lundqvist, U., Robèrt, K. H., and Wackernagel, M. (1999). *Int. J. Sust. Dev. World Ecol.* 6, 17.
- Kjellén, B., Näslund, S., and Olsson, K. (2002). "Forum on Science, Technology and Innovation for Sustainable Development, Johannesburg, 2002, Swedish Ministry of Environment, Internal Report" [in Swedish].
- Meyer, A. M., *et al.* (1995). *Nature* 378, 433.
- Råde, I., and Andersson, B. A. (2001). *J. Power Sources* 93, 55.
- Rees, W. E., and Wackernagel, M. (1994). In "Investing in Natural Capital: The Ecological Economics Approach to Sustainability" (A. M. Jansson, M. Hammer, C. Folke, and R. Costanza, Eds.), p. 362. Island Press, Washington DC.
- Robèrt, K. H., Holmberg, J., and von Weizsäcker, E. U. (2000). *Greener Manage. Int.* 31, 25.
- von Weizsäcker, E. U., Lovins, A. B., and Lovins, L. H. (1997). "The Factor four—Doubling Wealth, Halving Resource Use." Earthscan, London.
- Wall, G. (1985). "Exergy—A Useful Concept," Ph.D. thesis. Department of Physical Resource Theory, Chalmers, Göteborg, Sweden.

Plastics Production and Energy

MARTIN PATEL and NITIN MUTHA

Utrecht University
Utrecht, The Netherlands



-
1. Introduction
 2. Method and Scope
 3. Results
 4. Discussion
 5. Conclusion

Glossary

bio-based plastics Plastics that are fully or partially produced from biomass-derived feedstocks (renewable raw materials).

cradle-to-factory gate System covering all process steps from the extraction of resources to their transformation (e.g., in refineries) and the conversion to monomers and polymers.

cradle-to-grave Same as cradle-to-factory gate but including also all activities related to waste management (due to the diversity of application areas, the use phase is often excluded in comparisons for polymers).

finite energy Nonrenewable energy sources (i.e., mainly fossil and nuclear fuels).

petrochemical plastics Plastics that are produced from petrochemical feedstocks.

plastics Macromolecular materials produced from petrochemical feedstocks or renewable raw materials by means of polymerization.

Worldwide, the production of polymers totaled approximately 180 million tonnes in 2001. Of all man-made bulk materials, the total energy input for plastics production is estimated to be fifth after that for cement, steel, paper/board, and bricks/tiles. In the past few decades, plastics production outpaced that of all other bulk materials, with the average worldwide growth rate for plastics 5.8% per year between 1990 and 2000. During the same period, the growth rate of plastics production exceeded gross domestic product growth [3.6% per annum (PA)] by more than 2% PA. Plastics enjoy high growth rates both in low-

income and high-income countries (e.g., approximately 40% PA in Western Europe versus approximately 300% PA in India between 1990 and 2000). Given the very low per capita consumption rates in developing countries (India, 3.5 kg/capita; China, 12 kg/capita; world, 24 kg/capita), plastics production can be expected to continue to grow considerably in the next decades.

1. INTRODUCTION

Among the bulk plastics, use of polyolefins has increased most in the past decade, with particularly high growth rates for polypropylene (PP) and linear low-density polyethylene (LLDPE). Polystyrene (PS) and polyvinyl chloride (PVC), on the other hand, have grown at a rate approximately the world average. Table I provides an overview of products that are typically manufactured from these and other types of polymers.

The overwhelming share of all plastics is produced from petrochemical feedstocks (more than 99.9%). This is the consequence of the growth of the petrochemical industry in the early 20th century, whereas the first man-made polymers were derived from biomass resources. However, since the 1980s and 1990s, increased use of bio-based polymers has been observed in certain application areas. In 1998, production of bio-based polymers in the European Union (EU) was approximately 25,000 tonnes. The total market is projected to grow autonomously to approximately 500,000 tonnes until 2010 [28% per annum (PA), 1.5% market share] and to approximately 1 million tonnes if supportive policies and measures are implemented (36% PA, 3% market share). So far, there are no policies and measures at the EU level, whereas national legislation provides advantages for bio-based polymers in certain areas (e.g., due to the packaging ordinance in Germany).

TABLE 1
Overview of Typical Applications by Types of Polymers

Polymer	Applications
Epoxy resin	Industrial floorings, foams, fiber-reinforced plastics
High-density Polyethylene	Moulded products, heavy-duty films
Linear low-density polyethylene	Films (food packaging), rotational moulded products (e.g., tanks)
Low-density polyethylene	Food packaging
Melamine formaldehyde resin	Decorative laminates, lacquers
Phenol formaldehyde resin	Laminates for printed circuit boards
Polyacrylonitrile	Fibers
Polyamide (nylon)	Fibers, plastic gears, tire cord
Polycarbonate	Safety goggles, safety transparent shields, machinery housings
Polyethylene terephthalate	Blow-moulded bottles
Polypropylene	Injection-moulded products (e.g., technical parts for washing machine), films (e.g., for cigarette wrappers), monofilaments, ropes
Polystyrene	Moulded containers, lids, jars, bottles, foamed plastics
Polyurethane	Foams, coatings, adhesives and elastomers
Polyvinyl chloride	Extrusion products (wire and cable, conduits, profiles)
Unsaturated polyesteralkyd resin	Fibers, coatings, adhesive
Urea formaldehyde resin	Paper coating, adhesive

Most of the bio-based plastics that are currently on the market are biodegradable, but there are also increasingly more bio-based materials that are not biodegradable (e.g., composites of natural fibers and petrochemical polymers).

Starch polymers represent by far the largest group of commercially available bio-based plastics. Other prominent bio-based plastics are polylactic acid (PLA), polytrimethylene terephthalate, and polyhydroxyalkanoates (PHA).

Due their importance, starch polymers, PLA, PHA, and natural fiber composites, which are all manufactured at an industrial scale, are discussed in this article in addition to conventional petrochemical plastics. The first three materials are biodegradable; this is not the case for the natural fiber composites and the petrochemical plastics covered.

Given the dynamic developments discussed previously and the variety of plastics on the market, it is of great relevance to provide an overview of the energy use related to plastics. If the total life cycle of plastics is taken into account—comprising all process steps from the extraction of resources to polymer production and plastics processing and waste management—plastics account for approximately 4% of the total primary energy use in high-income countries (4.2–4.4% in Western Europe; data exclude the use phase). More than 90% of this is required for all processes related to plastics production, on which we will focus in this article (approximately 4.0% of total fossil fuel use in Western Europe). In the future, improved waste management, including recycling, reuse, and advanced energy recovery, may contribute substantially to materials and energy savings. These options and the attendant potential for energy saving are not discussed in this article.

2. METHOD AND SCOPE

This article is based on a review of energy analyses and life cycle studies on petrochemical and bio-based plastics. The studies reviewed basically follow the principles of life cycle assessment (LCA), which is a standardized method to quantify environmental impacts. In practically all LCA studies, energy use is included in the set of indicators, which typically also includes the release of greenhouse gases, the acidification and eutrophication potentials, waste generation, and other environmental impacts.

As in most LCA studies, we focus on the use of finite (i.e., nonrenewable) energy. This comprises fossil fuels and, to a lesser extent, nuclear fuels. Large amounts of renewable resources (especially biomass derivatives) are used for the production of bio-based plastics, but the quantities used are not discussed in this article. We focus on finite energy because of the limited resource perspective, the high dependence on foreign supplies, and the generally higher environmental impacts than in the case of renewable energy.

As indicated previously, we focus on energy use for plastics production, also referred to as cradle-to-factory gate analyses. This type of analysis includes all processes, including the extraction of energy resources, their transformation (mainly in refineries and power stations), the conversion of the feedstocks to monomers and polymers, and the manufacture of the required auxiliaries (Fig. 1, gray box). Processing of polymers to final products, their use phase, and

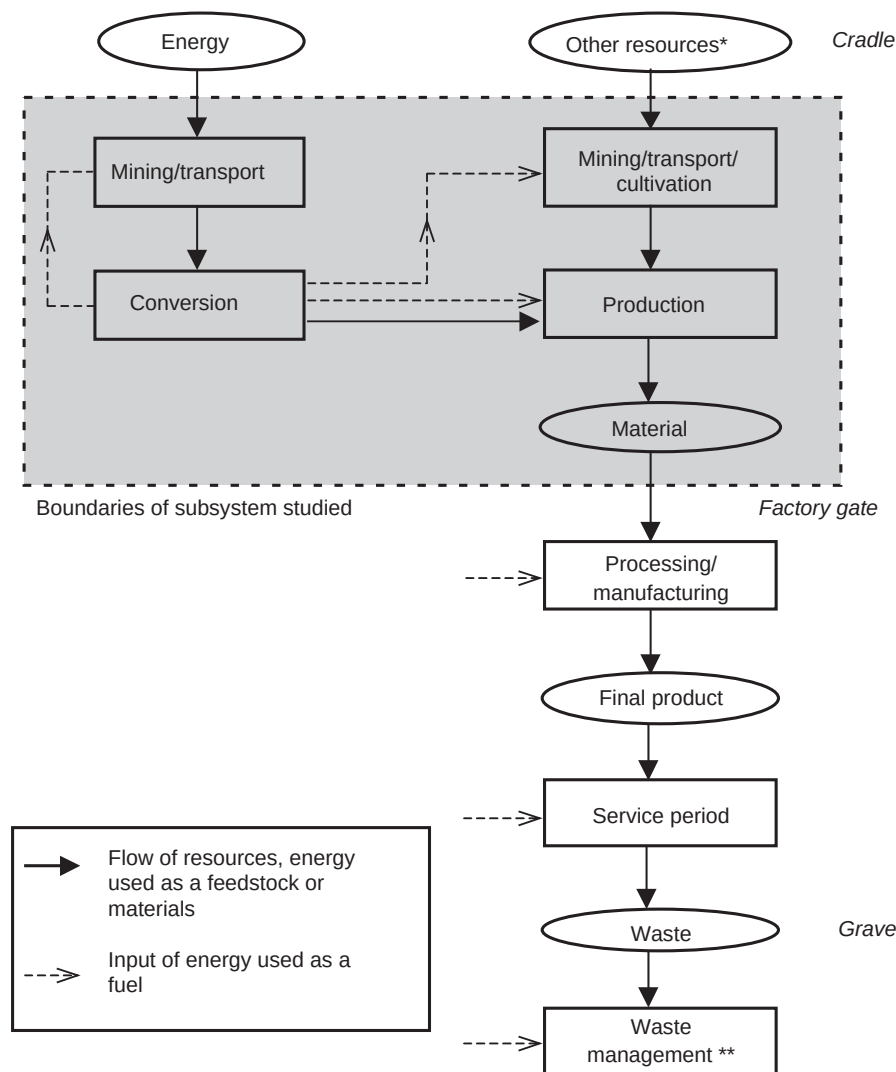


FIGURE 1 Overview of the entire system and the subsystem discussed in this article (gray box). *Energy analyses do not account for the consumption of resources other than energy, but they do include the energy requirements to process these “other resources.” **Landfilling, incineration, and recycling (not shown).

the waste management stage are not included in the system boundaries (in contrast, studies that do include the latter are referred to as “cradle-to-grave analyses”).

Fossil fuels are used both as a feedstock for chemical reactions and to cover the energy requirements of the production processes and transportation (Fig. 1). In the case of bio-based plastics, the biomass-derived inputs are almost exclusively used as a feedstock but efforts are being made to integrate waste and by-products in the energy supply.

As a basis for all comparisons, 1 tonne of plastics in the form of pellets (granules) has been chosen as the functional unit. This is a convenient basis since it

allows the comparison across types of polymers, but it also has some drawbacks.

3. RESULTS

A distinction can be made between the use of energy resources as feedstocks and as process energy (in the form of electricity, steam, or direct fuel use). Reporting of feedstock energy is complicated by the fact that in some processes, part of the feedstock is used to fuel the process (e.g., in steam crackers for olefin production and in ammonia production). As a consequence, feedstock energy use should be

defined precisely in every study, but this is rarely done (if feedstock and process energy are distinguished at all). For this reason, we do not distinguish between these two types of energy use and exclusively report total energy use (cradle-to-factory gate energy).

However, there is a relatively safe way to address the issue by making use of information on the heating value of plastics. The heating value represents the amount of feedstock energy that is physically fixed in the plastics (Table II). Hence, the remainder (i.e., the difference between cradle-to-factory gate energy and the heating value) can be defined as the process energy consumed in the various production steps.

In the following sections, we discuss energy requirements for the production of petrochemical plastics and bio-based plastics.

3.1 Petrochemical Plastics

Table III provides an overview of cradle-to-factory gate energy requirements of 21 petrochemical primary plastics (pellets and granules; compare Table I for typical applications). The studies identified differ with regard to the geographic scope (only high-income countries are represented) and the reference year (the oldest source dates back to 1980). There is also a major difference concerning the coverage of types of plastics: Whereas some of the sources speci-

fically focus on plastics production and/or recycling, others cover the most important building materials or even all major bulk materials.

A key source used frequently in Europe (also in LCA software tools) is the Eco-Profiles published by the Association of Plastics Manufacturers in Europe (APME). Table III shows two data sets from APME, one of which was published in 1999 and the other pre-1999. Other widely disseminated data sets are those released by the U.S. Office of Industrial Technology (OIT). It is interesting to note that substantial corrections have been made in the revised APME data set and that the values in the two data sets published by the OIT differ considerably (Table III). The difference in values across the other sources is even larger.

We calculated the standard deviation for those plastics for which at least 7 independent sources are available (polyethylene, polyvinyl chloride, polystyrene, polyethylene terephthalate, polyurethane, and nylon 6,6) and determined values ranging between 18 and 23% of the arithmetic mean. Assuming a normal distribution, the cradle-to-factory gate energy requirements have a probability of 68% within a range of $\pm 23\%$. For the remaining 14 plastics, there are less than 7 independent sources. The data from approximately half of these fall into the uncertainty range established for the 7 plastics, whereas the spread is larger for the others. Due to the low number of independent sources, quantification of the uncertainty range for the 14 plastics is hardly possible.

Especially for sources covering a larger number of materials, it is possible to make a statement about the relative position of the values: The data reported by APME (both data sets) are in the medium range, whereas the values published by Kindler and Nikles are partly in the medium range and partly on the higher side (PE is an exception since it is on the lower side). Data from studies by Alcorn and by Shutov are on the higher side, whereas those published by Patel and by the Plastic Waste Management Institute, Japan, are in most cases lower than those in the other sources (with the exception of nylon 6,6).

For the data published by the APME and by Patel, relatively detailed information is available about the assumptions and system boundaries that allow discussion of the reasons for the discrepancy. Except for ammonia, nylon-6,6, polyethylene, and polyvinyl chloride, the energy requirements determined for Germany are generally lower than those determined by the APME for Western Europe. The main reasons are as follows:

TABLE II

Heating Value of Bio-Based and Petrochemical Polymers^a

Polymer	Lower heating value, GJ/tonne (dry matter)
Starch polymers	13.6
Polyhydroxybutyrate	22.0
Polyhydroxyvalerate	25.0
Polylactic acid	17.9
Lignin (picea abies)	24.2
China reed	18.0
Flax	16.3
Hemp	17.4
Kenaf	16.5
Polyethylene	43.3
Polystyrene	39.4
Polyethylene terephthalate	22.1
Polyvinyl chloride	17.9

^aHeating values calculated according to Boie, compare Reimann and Hämmerli, 1995.

- The data for Germany refer to the net calorific value (NCV; lower heating value), whereas the APME publications use gross calorific values (GCVs; higher heating value). The difference is within 10%.

- The APME data are based on surveys for all of Western Europe. They may therefore include some smaller, older plants that are less efficient than the average of all units in Germany in the mid-1990s.

- The APME data accounts for the various shares of purchased versus autogenerated electricity, including the respective efficiencies for the sites and the public grids in the various countries covered. In contrast, exclusive autogeneration of power in efficient combined heat and power (CHP) plants has been assumed for plastics production in Germany.

- The efficiencies for steam generation assumed in the APME publications are smaller than the average assumed in the calculations for Germany.

There are also several aspects that have not been taken into account in the calculations for Germany but that are covered by the APME data:

- For specific processes, data were only available for the selectivities and not for the yields of the products. In these cases, the calculations are based on selectivities instead of yields, thus assuming a conversion of 100% (yield = selectivity $\times$ conversion). Consequently, material efficiency is slightly overestimated and therefore the results for the energy requirements in Germany are somewhat underestimated.

- The assumed losses in the steam system (7.5%) in plastics production in Germany may be too low, as indicated by the literature data on the losses related to degasification of the boiler feed water, elutriation of the boiler, and noncovered recycling of condensate indicate. On the other hand, the German Association of Chemical Industry estimates that the average steam losses are 2.5–4% (G. Thomas, personal communication, 1999).

- Facilities outside the battery limits (e.g., flue gas scrubbers and sewage treatment plants) are not included in the system boundaries and, consequently, they are not included in the data for Germany.

- Ancillary materials (e.g., packaging materials or lubricants), catalysts, and other materials used in small amounts have not been taken into account in the calculations for Germany.

- It is possible that the operating rates assumed in the calculations for Germany are higher than in reality. Moreover, the process data used for Germany refer to continuous operation with no maintenance

intervals (i.e., they exclude energy-intensive reactor start-ups and shutdowns).

- Transport energy used in the extraction and supply of resources has been accounted for (Fig. 1). However, due to the lack of detailed data, transport between all the subsequent conversion steps was not included in the calculations for Germany. Since the majority of the products studied were transported in pipelines, which does not consume a large amount of energy, and since transport energy generally only represents a small share of the total energy consumption, the errors resulting from this procedure can be assumed to be negligible.

To summarize, it may well be that the data for Germany underestimate the real energy requirements. However, the different choices made regarding the aspects listed previously and the differences in the years and regions analyzed largely explain the difference compared to the APME data. The results for Germany are therefore considered to be representative estimates for this country in the mid-1990s.

Despite the ranges in values in Table III, most studies confirm that relatively little energy is required to manufacture the bulk plastics PE and PVC. In contrast, the energy required for the manufacture of engineering and other plastics can be more than twice as high.

3.2 Bio-Based Plastics

Contrary to petrochemical plastics, the feedstocks used to produce bio-based plastics are derived from agricultural crops (e.g., maize, wheat, or sugar beet) or from suitable waste streams (e.g., from the dairy industry). Regarding petrochemical plastics, processing, cultivation, and transportation almost exclusively rely on finite energy sources, particularly fossil fuels. For bio-based plastics, the cultivation of crops (or the supply of suitable waste streams) results in an energy demand that is absent in the case of petrochemical plastics. Moreover, feedstock preparation and product separation may be clearly higher in the case of bio-based plastics due to relatively low yields and high water content in the raw material and product streams. It is thus not obvious from the outset that cradle-to-factory gate energy requirements for bio-based plastics are lower compared to those for petrochemical plastics.

However, Table IV shows that this is the case for nearly all of the bio-based plastics studied: thermoplastic starch (TPS), starch copolymers [blends of TPS with polycaprolactone (PCL) or polyvinyl

TABLE III

Energy Requirements for the Production of Petrochemical Plastics^a

Plastic	Total primary energy use (cradle-to-factory gate) (GJ/t)											
	APME (1999) ^b Western Europe	Pre-1999 APMEb and BUWAL Western Europe	Curlee and Das (1998) USA	Howard (1995) UK	Kindler and Nikles (1980) Germany	Alcorn (1998) New Zealand	OIT (2002) USA	OIT chemical industry profile reports USA	Patel (1999) Germany	PWMI Japan (1999) Japan	Shutov (1999) USA	Worrell <i>et al.</i> (1994) The Netherlands
Epoxy resin	140.7	141.7 ^c							107.1			
Melamine resin									79.9			
Nylon-6					156			124.3	122.7			
Nylon-6,6	142.3	143.6 ^d			154		241.7	135.1– 157.4	163.0		147.6	
Phenol formaldehyde resin					82				60.0		89.2	
Polyacrylonitrile									69.3			
Polyacrylate	PMMA, 102.3	PMMA, 111.7 ^e			PMMA, 91				82.6 ^f			
Polycarbonate	116.8	116.3 ^g			107				80.3		158.5	
Polyethylene (PE)				85				74.1– 80.5	64.6			
<i>Polyethylene, high- density</i>	79.9	81.0 ^b	84.8		70–72	103	131.5			65.7	98.1	
<i>Polyethylene, low- density</i>	80.6	88.6 ^b	89.5		69	103	136.0			69.2 ⁱ	103.2	67.8
<i>Polyethylene, linear low-density</i>	72.3	83.0 ^b					121.1					
Polyethylene terephthalate (PET)	76.2– 77.5 ^j	81.7–83.8 ⁱ	113.2		84–108	106			59.4	60.0 ^k	106.4	78.2
Polypropylene			79.5		73	64 ^l	126.0	54.5			95.3	63.2
Polystyrene	86.7 ^m	86.3 ⁿ	79.7	96	80–82			100.0– 104.5	70.8	68.5	117.1	82.7
Polyurethane (PUR)	104.4– 104.9 ^o	104.4– 104.9 ^p		65	98	74			75.5 ^q		73.7	
Polyvinyl acetate		66.8 ^r							58.2			
Polyvinyl chloride	56.6 ^s	56.7 ^t	59.5	59	53	70	34.6	44.8	53.2	45.4	79.0	52.4

Synthetic rubber			64.4 ^u	SBR, 77.5
Unsaturated polyester/ alkyd resins	78	93.7– 100.4	64.5 ^v	
Urea resin			50.3	

^a Cradle-to-factory gate nonrenewable energy including both feedstock energy and process energy. Abbreviations used: APME, Association of Plastics Manufacturers in Europe; BUWAL, Bundesamt für Umwelt, Wald und Landschaft (Swiss Agency for the Environment, Forests, and Landscape); OIT, Office of Industrial Technologies; PWMI, Plastic Waste Management Institute; PMMA.

^b APME also provides data on renewable energy that are outside the scope of this article (see text). However, since the share of renewables is very low (mostly 0.3–0.4%), these quantities have not been deducted in this column. This has the advantage that the data listed here can be quickly compared with the original source without further calculations.

^c APME Eco-profiles, Report 12 (1997).

^d APME Eco-profiles, Report 15 (1997).

^e APME Eco-profiles, Report 14 (1997).

^f Modeled as polymethyl acrylate.

^g APME Eco-profiles, Report 13 (1997).

^h APME Eco-profiles, Report 3 (1993).

ⁱ LLDPE included.

^j Lower value for amorphous resin, higher value for bottle grade PET.

^k Bottle-grade PE.

^l In some tables of the same source reported to be 95.4 GJ/t.

^m Figure for GPPS (general purpose polystyrene). For comparison, HIPS (high impact polystyrene), 92.2 GJ/t.

ⁿ Figure for GPPS. For comparison, HIPS, 90.7 GJ/t.

^o Lower value for flexible PUR foam, higher value for rigid PUR.

^p Lower value for flexible PUR foam, higher value for rigid PUR. APME Eco-profiles, Report 9 (1997).

^q 50% rigid polyurethane, 50% flexible polyurethane.

^r APME Eco-profiles, Report 6 (1994).

^s Figure refers to suspension PVC, which is the general-purpose grade and is used for most mass applications. For comparison, bulk PVC (used for specific types of hard sheets and bottles); 63.1 GJ/t; emulsion PVC, 67.7 GJ/t.

^t APME Eco-profiles, Report 6 (1998). Figure refers to suspension PVC, which is the general-purpose grade and is used for most mass applications. For comparison; bulk PVC (used for specific types of hard sheets and bottles); 63.2 GJ/t; emulsion PVC, 68.0 GJ/t.

^u Weighted average of the German production of SBR (styrene butadiene rubber), BR (butyl rubber), NBR (nitrile butadiene rubber), and CR (chloroprene rubber) in the mid-1990s.

^v Modeled as resin synthesized from phthalic anhydride (1 mol) and propylene glycol (2 mol).

TABLE IV

Energy Requirements for the Production of Bio-Based Plastics^a

Type of plastic	Cradle-to-gate finite energy use (GJ/t) ^b	Reference
Bio-based plastics (pellets)		
TPS	25.4	Patel <i>et al.</i> (1999)
TPS	25.5	Dinkel <i>et al.</i> (1996)
TPS	25.4	Dinkel <i>et al.</i> (1996)
TPS (maize starch + 5.4% maize grit + 12.7% PVOH)	18.9	Würdinger <i>et al.</i> (2002)
TPS + 15% PVOH	24.9	Patel <i>et al.</i> (1999)
TPS + 52.5% PCL	48.3	Patel <i>et al.</i> (1999)
TPS + 60% PCL	52.3	Patel <i>et al.</i> (1999)
Starch copolymer foam grade	32.4–36.5	Estermann <i>et al.</i> (2000)
Starch copolymer film grade	53.5	Estermann and Schwarzwälder (1998)
PLA	54	Vink (2002)
PHA by fermentation	81	Gerngross and Slater (2000)
PHA, various processes	66–573	Heyde (1998)
Natural fiber composites		
ABS copolymer ^c	117	Wötzel <i>et al.</i> (1999)
Hemp fiber/epoxy composite ^c	89	Wötzel <i>et al.</i> (1999)
Glass-fiber reinforced polypropylene ^d	93	Corbière-Nicollier <i>et al.</i> (2001)
China reed reinforced polypropylene ^d	61	Corbière-Nicollier <i>et al.</i> (2001)
Petrochemical polymers		
HDPE	79.9	APME (Boustead) (1999)
LDPE	80.6	APME (Boustead) (1999)
LLDPE	72.3	APME (Boustead) (1999)
PET (bottle grade)	77	APME (Boustead) (1999)
PS (general purpose)	87	APME (Boustead) (1999)
Expanded PS	84	APME (Boustead) (1999)
Petrochemical copolymers		
PCL	83	Estermann and Schwarzwälder (1998)
PCL	77	Kopf (1999)
PVOH	102	Estermann <i>et al.</i> (2000)
PVOH	58	Kopf (1999)

^a Cradle-to-factory gate nonrenewable energy including both feedstock energy and process energy. Complete bibliographic information on the references cited in this table can be found in Patel *et al.* (2003). Abbreviations used: TPS, thermoplastic starch; PVOH, polyvinyl alcohol; PCL, polycaprolactone; PLA, polylactides; PHA, polyhydroxyalkanoates; ABS, acrylonitrile, butadiene, styrene; HDPE, high-density polyethylene; LDPE, low-density polyethylene; LLDPE, linear low-density polyethylene; PET, polyethylene terephthalate; PS, polystyrene; APME, Association of Plastics Manufacturers in Europe.

^b Total of process energy and feedstock energy. Finite (nonrenewable energy) only; i.e., total of fossil and nuclear energy. In the cradle-to-factory gate concept, the downstream system boundary coincides with the output of the polymer or the end product. Hence, no credits are ascribed to potential benefits in the use phase to valuable by-products from waste management (steam, electricity, and secondary materials).

^c Application studied: Interior side panel for passenger car; ABS stands for acrylonitrile-butadiene-styrene.

^d Application studied: Transport pallet.

alcohol (PVOH), polylactides (PLA), and the natural fiber composites. For polyhydroxyalkanoates (PHA), the energy use strongly depends on the process and feedstocks used and ranges between values that are somewhat lower to those that are clearly higher than for petrochemical plastics. To facilitate direct com-

parison, Table IV also provides the energy requirements for the most important petrochemical polymers and for PCL and PVOH, which are both used as copolymers for starch plastics. Life cycle practitioners consider the data for PCL and PVOH to be subject to major uncertainties.

Most of the materials listed in Table IV can be produced according to the current state-of-the-art. Future options (e.g., the production of PHA in crops and the medium- and long-term production processes for PLAs) are not listed in Table IV. As an exception, technical problems still need to be resolved for pallets reinforced with china reed (material stiffness).

It can be concluded from Table IV that the cradle-to-factory gate energy requirements of pure TPS are 65–75% below those for PE. For starch polymers with high amounts of petrochemical polymers, the energy requirements are approximately 30% lower than those for polyolefins. However, the applications for pure starch polymers and blends with small amounts of petrochemical copolymers are limited due to inferior material properties (especially sensitivity to moisture). With larger amounts of petrochemical copolymers, polymers with wider application profiles can be produced. This allows to obtain higher energy savings in absolute terms while the marginal energy savings decrease with higher shares of petrochemical copolymers.

For PLA, data are only available from a source (Cargill Dow), according to which cradle-to-factory gate energy requirements are 20–30% less than those for PE.

The results reported for PHA vary greatly. Cradle-to-factory gate energy requirements in the best case (66.1 GJ/t) are 10–20% lower than those for PE. The upper range in Table IV refers to substrate supply from fossil-based feedstocks (e.g., methanol).

It can be concluded that of all bio-based plastics studied, starch polymers perform best in environmental terms under the current state-of-the-art, with some differences among the various types of starch polymers. Compared to starch polymers, the environmental benefits seem to be smaller for PLA. For PHA, the environmental advantage seems to be very small compared to that for conventional plastics in the best case; in the worst case, it can be substantially higher. For both PLA and PHA, the production method and the scale of production can influence decisively the cradle-to-factory gate energy requirements.

The extent to which natural fibers (e.g., hemp and flax) can replace fiberglass (which is heavy and energy intensive to produce) determines the net environmental benefits. Interesting possibilities are also available.

Since most of the data in Table IV refer to the current state-of-the-art, technological progress, improved process integration, and various other possi-

bilities for optimization will likely provide more favorable results for bioplastics in the future.

4. DISCUSSION

In the preceding section, all data were compared on the basis of 1 tonne of plastics. Although this is a convenient choice, it suffers from the drawback of not taking differences in material properties and the resulting parameters at the end use level into account. For cradle-to-factory gate analyses, this mainly concerns (i) materials processing, in which the amount of material required to manufacture a certain end product and the energy use in the processing step may differ depending on the type of plastic (e.g., bio-based vs petrochemical plastics), and (ii) transportation energy, which can be substantial for end products with a low bulk density such as packaging chips (so-called loose fill).

On the other hand, these aspects can only be fully taken into account by analyses dedicated to specific end products (e.g., plastic bag), whereas the choice of one mass unit of polymers as a functional unit allows a preliminary conclusion about the advantages or disadvantages in terms of energy use. If, for example, the energy requirements for a given polymer are relatively high at the level of primary materials (pellets), there is a good chance that this will also be true at the end product level.

The studies reviewed in this article reflect the situation in different countries. The types of feedstocks used and the processes applied can vary from country to country or from continent to continent. For example, olefin production in steam crackers is mainly based on naphtha in Europe and Japan, whereas in the United States natural gas-based ethane crackers dominate. Differences in processes and feedstocks may result in different yields and energy efficiencies. Due to the competition in the sector, the differences in efficiencies of the production processes can be assumed to be relatively low across high-income countries. This may be less the case for power production: Here, the requirements of (finite) energy resources depend to a large extent on the type of fuel used (e.g., coal with relatively low efficiencies vs natural gas with high ones), the share of CHP, and the percentage of renewable energy (generally very low but high in Norway and Switzerland).

Direct comparison of energy data is complicated by the fact that both the NCV and the GCV are used in this type of analysis but that the choice is not always clearly reported. This also concerns the

inclusion or exclusion of ancillary materials, catalysts, and additional facilities such as sewage treatment plants that generally have a much lower impact on the results. This can also be assumed for different reference years as long as they do not differ by more than a decade. Larger discrepancies among the studies may accrue from different allocation procedures for processes with larger amounts of coproducts. Also in these cases, the concrete choices are hardly ever explained in detail.

An attempt should be made to consider these aspects when using cradle-to-factory gate energy data for further analysis and when interpreting the results and drawing conclusions. Moreover, whether the waste management system should also be included in the system boundaries (cradle-to-grave system) should be carefully considered.

5. CONCLUSION

By combining the cradle-to-factory gate energy data with current production data, the global primary energy use for the production of polymers is estimated to be approximately 15,000 PJ or 3.6% of the global primary energy use. Although the uncertainty of the production data is unknown, the specific energy requirements for industrialized countries have been estimated with a probability of 68% within a range of $\pm 23\%$. Since the energy efficiency in developing countries is generally lower, the global value can be considered a conservative estimate. In view of the high current and anticipated growth of plastics production, the attendant energy use is very likely to increase substantially. Basically, four strategies can be distinguished to minimize this growth:

1. Energy efficiency: The most straightforward strategy is increased energy efficiency in the petrochemical processes and in polymerization and increased use of production routes and intermediates with low overall energy requirements. For example, PVC, which was formerly manufactured from energy-intensive acetylene, is currently mainly being produced from ethylene. Research and development efforts are directed toward the use of ethane as a feedstock.

2. Material substitution: With advances in polymer science, it may become increasingly possible to replace energy-intensive engineering plastics (such as polyamides) with modified bulk polymers (e.g., polyolefins), which are more energy efficient to produce.

3. Waste management: Various types of recycling, the reuse of plastic components, and energy recovery

can contribute to lower life cycle energy requirements and environmental impacts.

4. Bio-based plastics: As apparent from Section III, the substitution of bio-based raw materials for petrochemical feedstocks may provide an interesting potential for energy saving (Table IV).

Energy analysis can be a useful tool for establishing the relative benefits of specific measures of these strategies.

When conducting an energy analysis, the researcher is often confronted with the fact that different sources provide different values for a certain product and that the deviation between them can be substantial (Table III). This is generally due to different system boundaries (e.g., concerning the years and geographical regions studied and the technologies and processes covered). It is a prerequisite for a reliable analysis that the cradle-to-factory gate data used have been determined in a consistent manner. In contrast, comparative assessments that use energy data originating from various sources can easily result in distorted conclusions given the large data ranges shown in Table III (differences of a factor of two in extreme cases). These data ranges and the corrections made by the APME indicate that considerable uncertainties continue to exist despite the fact that petrochemical polymers are manufactured by use of mature technologies that are applied globally with only limited variations. On the other hand, there is good agreement concerning the relative position of the various plastics.

As a consequence, further analyses on cradle-to-factory energy requirements for plastics should be conducted. For the time being, it is suggested to use, wherever available, data sets that are valid for the country (e.g., Germany) or region (Western Europe) under investigation. In order to check the reliability of the results, it is recommended to conduct sensitivity analyses using other data sets that have been determined in a consistent manner for a different region than the one investigated. Alternatively, calculations on error propagation can be conducted.

SEE ALSO THE FOLLOWING ARTICLES

Aluminum Production and Energy • Cement and Energy • Glass and Energy • Material Efficiency and Energy Use • Material Use in Automobiles • Steel Production and Energy

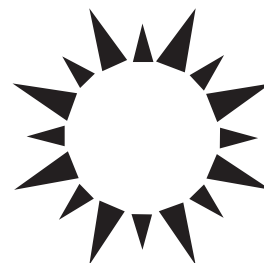
Further Reading

- Alcorn, A. (1998). "Embodied Energy Coefficients of Building Materials," 3rd ed. Victoria University of Wellington, Centre for Building Performance Research, Wellington, New Zealand. [http://www.arch.vuw.ac.nz/cbpr/index_embodied_energy.html]
- Association of Plastics Manufacturers in Europe (1993–1998). Eco-profiles of the European plastics industry (various reports on different types of plastics). Prepared by I. Boustead for the Association of Plastics Manufacturers in Europe, Brussels, Belgium.
- Association of Plastics Manufacturers in Europe (1999). "Eco-Profiles of Plastics and Related Intermediates." Association of Plastics Manufacturers in Europe, Brussels, Belgium. [<http://lca.apme.org>].
- Ayres, R. U. (1995). Life cycle analysis: A critique. *Resour. Conserv. Recycling* **14**, 199–223.
- Curlee, T. R., Das, S. (1998). The viability of recycling plastic by tertiary processes. *J. Solid Waste Technol. Management* **25**(3/4), 148–160.
- European Commission (2001). "ECCP (European Climate Change Programme)—Long Report." European Commission, Brussels, Belgium. [http://europa.eu.int/comm/environment/climat/eccp_longreport.0106.pdf].
- Howard, N. (1995). Environmental impact of building materials, Volume D, plastics and elastomers, Project Report 14. Construction Industry Research and Information Association, London.
- International Monetary Fund (2001). "World Economic Outlook." International Monetary Fund, Washington, DC.
- International Organisation for Standardisation (1997). "Environmental management—Life cycle assessment—Principles and framework, ISO 14 040 (DIN EN ISO 14 040)." Deutsches Institut für Normung, Berlin.
- International Organisation for Standardisation (1998). "Environmental management—Life cycle assessment—Goal and scope definition and inventory analysis, EN ISO 14 041." Deutsches Institut für Normung, Berlin.
- International Organisation for Standardisation (1999). "Environmental management—Life cycle assessment—Life cycle impact assessment, EN ISO 14 042 (draft)." Deutsches Institut für Normung, Berlin.
- International Organisation for Standardisation (1999). "Environmental management—Life cycle assessment—Life cycle interpretation, EN ISO 14 043 (draft)." Deutsches Institut für Normung, Berlin.
- Joosten, L. (2001). The industrial metabolism of plastics. Analysis of material flows, energy consumption and CO₂ emissions in the life cycle of plastics, PhD thesis. Utrecht University, Utrecht, The Netherlands.
- Kindler, H., and Nikles, A. (1980). Energieaufwand zur Herstellung von Werkstoffen—Berechnungsgrundsätze und Energieäquivalenzwerte von Kunststoffen. *Kunststoffe* **70**(12), 802–807.
- Office of Industrial Technologies (2000). "Energy and Environmental Profile of the U.S. Chemical Industry." U.S. Department of Energy, Washington, DC. [http://www.oit.doe.gov/chemicals/pdfs/profile_chap1.pdf].
- Office of Industrial Technologies (2002). "Energy, Environmental, and Economics (E3) Handbook." Energy Efficiency and Renewable Energy Network, U.S. Department of Energy, Washington, DC. [<http://www.oit.doe.gov/E3handbook>].
- Patel, M. (1999). Closing carbon cycles: Carbon use for materials in the context of resource efficiency and climate change, PhD thesis. Utrecht University, Utrecht, The Netherlands. [ISBN 90-73958-51-2; <http://www.library.uu.nl/digiarchief/dip/diss/1894529/inhoud.htm>]
- Patel, M., Bastioli, C., Marini, L., Würdinger, E. (2003). Life-cycle assessment of bio-based polymers and natural fibres. In: "Biopolymers" (A. Steinbüchel, Ed.), Vol. 10. Wiley-VCH, Weinheim, Germany.
- Plastic Waste Management Institute, Japan (1999). Report on survey of LCI data of petrochemical products. Overview—LDPE, HDPE, PP, PS, EPS, PVC and B-PET. Plastic Waste Management Institute, Japan, Tokyo. [Available from Hirokazu Hayashi, WIC, Senior Chief Researcher Office of Research & Information, Daiichi-Kaneichi Building, 3-32-3, Hongou Bunkyo-ku, Tokyo 113-0033, Japan]
- Reimann, D. O., and Hämmerli, H. (1995). "Verbrennungstechnik für Abfälle in Theorie und Praxis." Schriftenreihe Umweltschutz, Bamberg, Germany.
- Shutov, F. (1999). Effective energy and gas emission savings using plastics waste recycling technologies. Expert group meeting on industrial energy efficiency, cogeneration, and climate change. Environmental Sustainability Division/Kyoto Protocol Branch/United Nations Industrial Development Organisation. Vienna International Center, Vienna, Austria, December 2–3.
- VKE working group statistics and market research (German Association of Plastics Manufacturers) (2002). Plastics: Business data and charts. German Association of Plastics Manufacturers, Frankfurt, Germany. [http://www.vke.de/pdf-files/widat_engl.pdf]
- Worrell, E., van Heijningen, R. J. J., de Castro, J. F. M., Hazewinkel, J. H. O., Beer, J. G. de, Faaij, A. P. C., and Vringer, K. (1994). New gross energy-requirement figures for materials production. *Energy* **19**, 627–640.

Polar Regions, Impacts of Energy Development

BRUCE C. FORBES

University of Lapland
Rovaniemi, Finland



-
1. Regions of Impact
 2. Geography, Accessibility, and Economics of Energy Reserves
 3. Types of Impact
 4. Ecological Implications of Energy Development
 5. Sociocultural Implications of Energy Development

Glossary

corridor Linear infrastructure (e.g., road, railway, and power line) that cuts across otherwise intact vegetation/soil cover, such as closed boreal forest or tundra.

cumulative impacts Responses to disturbance, often relatively indirect, that differ from simple direct and additive effects by accumulating in space and time, sometimes in unforeseen ways.

disturbance Any sudden deviation (e.g., a reduction in species richness or biomass) from a reference state, which need not be static. A disturbance regime has four basic dimensions: frequency, intensity, scale (extent), and timing (season).

permafrost Perennially frozen ground; a substrate that is in a state at or below 0°C (32°F) for 2 or more years. It can be composed of soil, bedrock, water/ice, or any combination of these. Moving generally from north to south, the Arctic can be divided into regions of continuous, discontinuous, and sporadic permafrost.

rehabilitation Used as a broader term than revegetation to denote the establishment of productive and functioning systems and may include revegetation, erosion control, or fish transplanting efforts.

restoration A more restrictive term referring to the long-term process of reestablishing the original composition, structure, and function to a disturbed site.

revegetation Refers to the establishment of vegetation, regardless of composition, on a disturbed area. Can be either natural (unassisted) or assisted.

succession Generally involves changes in vegetation species composition or in the relative numbers of individual

plants through time at a given locale. It is often, but not always, precipitated by disturbance.

thermokarst The process by which characteristic landforms result from the melting of ice-rich permafrost.

tundra From the Finnish word *tunturi*, for sparse barrens near the timberline. Tundra is the Russian translation of this and refers more to the treeless zones beyond the latitudinal or arctic timberline as opposed to altitudinal timberline.

The polar regions hold enormous potential for the production of energy, mainly in the form of crude oil, natural gas and gas condensates, coal, and hydroelectric power. Some coal deposits in arctic Russia have been under industrial exploration and exploitation since the 1930s, and in other countries sources of hydropower and petroleum have also been under development for several decades. The Arctic contains by far the greater proportion of proven and economically viable deposits. In Antarctica, the Antarctic Treaty currently bans mining of any kind, although coal has been found in two regions and it has been suggested that rock formations there may well contain oil or gas. However, even if sizable reserves were discovered, and no exploratory drilling has been done, it is highly unlikely that the deposits could be recovered economically or safely. This article therefore focuses on the impacts of energy development in the Arctic, with special emphasis on oil and gas production, the main source of energy-derived revenue as well as a major cause of concern for environmental and sociocultural degradation. Examples are drawn from past and current energy developments from throughout the circumpolar North.

1. REGIONS OF IMPACT

The way we view energy development in the circumpolar North has partly to do with where we

come from. Americans tend to think of Alaska as the Arctic. Because of the construction of the Trans Alaska Pipeline (TAPS) in the 1970s, the oil spill from the Exxon Valdez in Prince William Sound in 1989, and the bitter fight over opening the Arctic National Wildlife Refuge (ANWR) to oil development during the past decade, many Americans are familiar with the image of Alaska as an important oil-producing state. In neighboring Canada, increasingly more Canadians are probably aware of the rekindled prospects for a major gas pipeline from the Beaufort Sea south via the Mackenzie River valley to central Alberta. Talk to British or Scandinavian nationals about oil and one will hear about the declining production from the North Sea. Also, whereas most Russians know of the political power wielded by the former state monopoly, Gazprom, relatively few outside of international business/finance circles recognize this company's name.

The fact is that arctic energy development has become an increasingly multinational domain driven by a high-tech, expensive, and growing network of extraction and transport infrastructure that spans virtually the entire Northern Hemisphere. Drivers in Europe fill up their cars with gasoline refined from crude oil pumped from beneath reindeer pastures in western Siberia. Homes in Chicago are heated in winter with natural gas piped halfway across Canada from the mountains of northern British Columbia. Village power plants in central Russia burn coal mined at a former gulag camp in the open tundra well north of the Arctic Circle in the faraway Komi Republic. New England buys surplus electricity generated from huge artificial lakes in the James Bay region of northern Quebec. A hydropowered smelter under construction in the highlands of eastern Iceland will soon be providing aluminum to the world market. A fleet of icebreakers and ice-strengthened tankers is being constructed to carry oil via Russia's northern sea route (NSR) to markets in Europe and the United States. In other words, arctic energy exemplifies the process of globalization.

To understand the impacts of energy development in the Arctic we must consider first the geography, accessibility, and economics of energy reserves within the Arctic rim countries: Russia, Finland, Sweden, Norway, Iceland, Greenland, Canada, and the United States. Nonrenewable and renewable resources vary widely in quantity and quality across the circumpolar North, as do the financial and human resources available to exploit them, not to mention the prevailing environmental conditions. Some countries, such as Finland and Sweden, have little or no

petroleum resources of their own and must import virtually all gasoline and fuel oils. They rely on nuclear plants, hydropower, coal, and even the burning of their plentiful woodchips to produce sufficient electricity for lighting and heating during the long, cold, and dark winters. Iceland has abundant geothermal energy available locally, but it also relies heavily on hydropower and imported petroleum to meet its energy needs. In contrast, Russia is considered the Saudi Arabia of natural gas and may one day be Europe's prime supplier.

2. GEOGRAPHY, ACCESSIBILITY, AND ECONOMICS OF ENERGY RESERVES

2.1 Alaska

Interest in petroleum development in Alaska dates to the early 20th century, when representatives of the U.S. Geological Survey sent to explore the North Slope provided reports about oil seeping out of the ground in places on the coastal tundra beyond the Brooks Range. In 1923, President Harding ordered that a 37,000 square mile area be set aside as National Petroleum Reserve No. 4. After the outbreak of World War II, the U.S. government became concerned about the domestic supply of oil and began seismically exploring the huge area. During the period 1944–1953, dozens of test wells were drilled in the zone of continuous permafrost. At the time, the area was extremely remote, and with no roads most travel across the tundra was necessarily by helicopter and tank-like off-road vehicles with metal tracks instead of wheels. The oil fields found were disappointingly small, although six sizable gas fields were discovered. It was not until January 1968, after 10 years of exploration, that Atlantic Richfield Company (ARCO), Humble Oil, and, later, British Petroleum (BP) tapped into a field estimated to contain at least 5–10 billion barrels (bbl). The strike at Prudhoe Bay turned out to be the largest ever made in North America and plans were drawn up for an 800-mile pipeline to extend from the North Slope to the ice-free port of Valdez.

Because the pipeline had to cross land that Alaska Natives claimed, the discovery at Prudhoe Bay immediately created pressure for a land claims settlement to access the oil fields. The Alaska Native Claims Settlement Act (ANCSA) of 1971 transferred title to 18 million ha of land and \$962.5 million to Alaska Natives. The land and money were not given

directly to individual Natives but, rather, the act divided the state into 12 regions and required that a state-chartered Native corporation be formed for each region with the provision that undeveloped Native corporation land was to be untaxed until 1991. There were problems of inequity because of the wide variation in exploitable natural resources. Natives who held claim over what became the North Slope Borough stood to profit from oil field leases far more than their neighbors. Although originally skeptical of Native claims regarding the critical importance of subsistence livelihoods, in 1987 some key amendments were made to ANCSA by which Congress recognized that the Native subsistence economy and culture are indeed significant factors in the management of Native lands and so attempted to provide for their protection. Subsistence provisions were also set forth in the Alaska National Interest Lands Conservation Act of 1980.

Since coming on line, the Prudhoe Bay/Kuparuk oil fields, via TAPS, have accounted for approximately 25% of the oil produced in the United States. Approximately 85% of the Alaskan state budget is derived from taxes and royalties collected on oil production. BP Exploration (Alaska), Incorporated and Phillips Alaska (formerly ARCO Alaska) are the two principal oil field operators and are both among the top 10 private employers in the state. At its peak in 1987, the North Slope produced an average of 1.61 million bbl of oil per day (bpd). In 2002, daily output declined to approximately 0.54 million bpd. Estimates of the recoverable reserves from the so-called 1002 area of ANWR vary widely from 4.3 to 30 billion bbl. This range has partly to do with economics because when market prices are less than \$35 per barrel, a larger volume of technically recoverable oil is considered to be economically recoverable. However, politics also plays a role in that different players prefer to publicize higher or lower estimates. In comparison, the great Prudhoe Bay field was originally estimated to hold approximately 9.2 billion bbl in total reserves, but this was later increased to 23 billion bbl of oil and 26 trillion ft^3 of natural gas.

2.2 Canada

Canada is the world's third-largest natural gas producer, after Russia and the United States, with approximately 50% of its 6.5 trillion ft^3 of annual production destined for the U.S. market. Overall, Canada supplies 16% of the gas consumed in the United States, double its share a decade ago. In the

Canadian Arctic, the major petroleum reserves lie in the Mackenzie River valley, delta, and Beaufort Sea region.

In the early 1900s, native Dene first led geologists to the site they called Le Gohlini, or "where the oil is," near Norman Wells on the lower Mackenzie River just north of the Arctic Circle. This led to the drilling of the first test well in 1919 by Imperial Oil, which confirmed the presence of rich oil fields. World War II led the U.S. Army to oversee the fast-track construction of the strategically located Canadian Oil Pipeline (Canol) almost 3000 km from Norman Wells to Whitehorse, Yukon. However, completion of the pipeline in less than 2 years coincided with the end of the war, and the project was quickly abandoned. In the early 1980s, Imperial Oil brought its Norman Wells operation to full capacity, and oil began flowing through the 870-km Norman Wells pipeline to Zama, Alberta. As of 2003, the pipeline averaged 10 million bbl per year (bpy) and was Canada's fourth-largest oil-producing field. Natural gas is also produced at Norman Wells, mainly for local use and for reinjection to enhance oil recovery.

Farther north, in the 1970s, some reserves of oil and gas were located in the high Arctic by Panarctic Oils. Feasibility studies revealed that a pipeline was not economically viable. Since the 1980s, there have been some shipments of oil eastward from Panarctic's Bent Horn field on Cameron Island and several experimental shipments of oil westward from Gulf Oil's Amauligak field in the Beaufort Sea. However, Canada's oil production is primarily south of the 60th parallel and gas comprises a much larger component of the nation's petroleum production. In recent years, exploration has been expanding off the coasts of Newfoundland and Labrador, where the federal and provincial governments are trying to replace the declining fishing industries with energy prospecting. Canada is also considering lifting a moratorium on oil and gas development off British Columbia.

The vast natural gas reserves of the Mackenzie River delta region were discovered in 1971. Whereas the Prudhoe Bay area of Alaska is estimated to contain as much as 100 trillion ft^3 of natural gas, an additional 64–73 trillion ft^3 lies in the combined onshore and offshore reserves of the Mackenzie delta and adjoining Canadian portion of the Beaufort Sea. A proposed 3500-mile Mackenzie River valley pipeline linking the American and Canadian fields would cost an estimated \$26 billion and take 7 years to build. A shorter pipeline exploiting only the Mackenzie delta field would cost approximately \$3 billion. The latter pipeline would have a capacity of approximately

1 billion cubic feet per day, equal to approximately 2% of U.S. consumption, whereas a pipeline joining northern Alaska would carry 4–5 billion ft³ per day.

For the longer pipeline, energy company engineers have been studying three possible routes from the Arctic Ocean to connect via Norman Wells to pipelines in Alberta. Proponents of an Alaskan route estimate that they could start pumping gas south in 6 years. They believe that they can work more quickly because permits and an international regulatory framework have existed since the late 1970s to place a gas pipeline parallel to the TAPS, which would then run along the Alaska–Canada Highway. Another plan would cut a new overland route across the Brooks Range and northern Yukon, and a third route would run up the Mackenzie River valley and have a submerged section running 15 ft below the floor of the Arctic Ocean for 300 miles to connect the Mackenzie River delta to Prudhoe Bay.

The latter route is the one officially favored by the government of the Northwest Territories, which has frequently been at odds with the government of the Yukon Territory over the choice of a pipeline route. The northern Yukon is estimated to have reserves of 3.3 trillion ft³ in the Eagle Plains and Peel Plateau and could potentially contribute 200 million ft³/day to a Mackenzie Valley pipeline. The main hub for regional development and administrative activities is Inuvik, Northwest Territories, with a population of 3000, making it Canada's largest community north of the Arctic Circle. Like Prudhoe Bay, it is accessible year-round by road from the south.

As of early 2003, the Aboriginal Pipeline Group was raising money to cover its share of the costs for preliminary design and regulatory work, regardless of the chosen route. Meanwhile, TransCanada PipeLines Ltd. was positioning itself to play the dominant role in delivering arctic gas to southern markets, although Calgary-based Enbridge was working to be heavily involved as well. In June 2003, a consortium consisting of TransCanada, Imperial Oil, and ExxonMobil's Canadian unit completed a partnership agreement with the Aboriginal Pipeline Group to build an 800-mile portion of the Mackenzie Valley pipeline.

2.3 Greenland

Greenland may be sitting on oil reserves comparable in size to those of the entire North Sea according to the Home Rule government's Bureau for Minerals and Petroleum (BMP). North Sea fields have produced approximately 29 billion bbl of petroleum to

date and proven remaining reserves stand at more than 15 billion bbl. A consortium led by Norway's Statoil and U.S. Phillips Petroleum has been awarded exploration licenses in waters off the southwest coast of Greenland. The BMP acknowledges that gigantic fields would be necessary to economically justify opening up the area.

2.4 Russia

Large-scale energy exploration and exploitation in arctic Russia have been going on for several decades. A coal mine at Vorkuta in northernmost Komi Republic began production as a gulag camp under Stalin in the late 1930s with a railway link to Moscow and other Russian regions. It grew quickly to a large complex of connected mining towns with an eventual peak population of 280,000 and coal production of 22 million tons per year by the late 1980s. Since 1991, the population has declined by approximately 30% and coal production has fallen well below levels from the mid-1970s. Petroleum has also long been under production in the Russian north, especially in the northern boreal and subarctic lowlands of western Siberia. Western companies began exploring projects in Russia even before the collapse of the Soviet Union, but more than a decade later they can point to few successes after combined investments of more than \$5 billion.

ExxonMobil, the world's largest oil firm, is focused mainly on its huge oil and gas project on Russia's far eastern island of Sakhalin and is therefore not directly involved in any truly arctic ventures. Shell, the world's second largest oil firm, signed a memorandum in the late 1990s to develop oil and gas condensate deposits at Gazprom's huge Zapolyarnoye gas field in western Siberia, but the deal was never finalized. BP, the world's third largest oil firm, has had extensive projects in the Russian north via its 50% interest in Tyumen Oil (TNK), which produces 1.2 million bpd and has reserves of 5.2 billion bbl. In June 2003, U.K. Prime Minister Blair and Russian President Putin jointly announced the creation of a new oil company, TNK-BP, which will lead to an additional \$6.15 billion investment by BP in Russia. TotalFinalElf, the world's fifth largest oil company, is developing a medium-sized oil field in Russia's north together with Norway's Norsk Hydro, Kharyaga, which produces 30,000 bpd and holds 710 million bbl of reserves. Total also acquired in 2000 the giant but remote Vankor field in eastern Siberia and has plans to develop with Gazprom the giant offshore arctic Shtokmanovskoye gas field. Other firms

actively engaged in Russian energy production and transport include ConocoPhillips, Finland's Fortum group, and Germany's Ruhrgas, which are involved in the financing, planning, and building of pipelines to European markets, mainly via the terminal facility at Primorsk, 100 miles (160 km) northwest of St. Petersburg.

In general, Russia faces enormous transport problems in getting its petroleum resources, for the most part deep in Siberia, to export markets. This makes scenarios for shipping oil to North America profitable only at times of high prices. Most of Russia's oil and gas goes to Europe. In the short term, U.S. oil companies could raise Russia's oil output by investing more in Russia. However, U.S. companies have been reluctant to invest there because of shifting tax codes and regulatory policies and have called for binding legislation before investing more money in the country. Specifically, they have been unwilling to do this without production sharing agreements, which Russia's Duma has been reluctant to grant foreign firms.

2.4.1 Murmansk Terminal

As the world's No. 2 oil exporter, Russia is seeking new markets outside Europe, whereas the United States, the world's largest oil consumer, wants to cut its reliance on the oil-rich but volatile Middle East. Under this scenario, Murmansk would become the main port for oil export to the United States. By 2010, Russian companies hope to cover 10–15% of crude oil consumption in the United States. Indeed, in November 2002 Presidents Bush and Putin pledged to strengthen energy ties between the two countries. The project is part of a blossoming relationship between the former Cold War rivals. However, Washington wants more than simply increased Russian oil imports out of its new energy relationship with Moscow. It is also keen to open Russia's vast oil reserves to U.S. investors.

In November 2002, Russia's top four oil companies—Lukoil, Yukos, TNK (since merged with BP), and Sibneft (since merged with Yukos)—announced their plan to sign a preliminary deal to build a 1-million bpd capacity oil-export terminal in Murmansk, Russia's only ice-free Arctic port. Oil to supply Western Europe and North America is expected to come primarily from the Timan–Pechora fields, at least initially. Timan–Pechora has one of the region's largest known reserves of oil, estimated to be approximately 126 billion bbl. Currently, ConocoPhillips produces approximately 25,000 bpd (10 million bpy) from its Polar Lights facility, a joint

venture with Lukoil along the border between the Nenets Autonomous Okrug (NAO) and Komi Republic. It also wants to develop a neighboring area, Northern Territories, believed to contain 4.25 billion bbl of crude under Russian standards, which tend to give larger estimates than Western rules. Transport would be either by ice-strengthened tankers loaded at the Lukoil-operated Varandei terminal in the NAO or via a proposed 80-million tons per year pipeline.

There are two options for the pipeline route. The first route is western Siberia–Ufa–Yaroslavl–Archangelsk–Murmansk. It is 2500 km long, including a short section which runs along the floor of the White Sea. The second route is western Siberia–Ukhta–Murmansk, which is 3600 km long. This route is fully on the ground and allows additional switching of the Kharyaginskoe field of Lukoil and TotalFina-Elf in the NAO to the transportation system. There are also proposals to modernize an existing pipeline, Ukhta–Yaroslavl, which transports oil from the Lukoil fields in the Komi Republic. The costs of the project are estimated to be \$3.4 billion for the first route and \$4.5 billion for the second route.

This is the first project of its kind and scale in Russia in which the companies plan to undertake the development essentially on their own. This would mean little or no participation by the state, which normally defines oil transportation via the trunk pipelines as a monopoly activity. To date, the fact that pipelines are built solely by state monopoly Transneft has allowed President Putin to maintain control over the up and coming petroleum oligarchs. However, most analysts insist that the development of such a grandiose project as construction of a large oil export facility in Murmansk, in combination with development and transport of the hydrocarbon reserves of Timan–Pechora, Yamal, and the Barents seabed, will not be possible without any state support. It is widely acknowledged that Russia has already reached an oil production bottleneck and cannot increase export capacity before expanding its transport infrastructure (ports, pipelines, and railways) and replacing a great deal of obsolete technology.

Russia's daily output is 8 million bpd, of which approximately 5.1 million bbl is exported. In comparison, U.S. domestic crude oil production averaged 5.8 million bpd in 2002, whereas oil imports averaged 9 million bpd for the same year, spiking to 9.4 million bpd in October. It is widely believed that the Russian arctic seabed may well become one of the main sources of oil in the 21st century. Moreover, oil resources of the North Sea are

steadily decreasing and it is estimated that reserves of the Norwegian sector are sufficient for only another 25 years and that Europe will soon import up to 85% of its energy supplies. Currently, Russia provides only approximately 20% of Europe's natural gas needs and 16% of its oil.

2.4.2 *Yamal*

In response to the situation described previously, Gazprom and Rosneft Oil Company and its affiliates for the exploitation of arctic oil and gas resources created Sevmorneftgaz as a joint enterprise. At the same time, many agreements between Russian and foreign consortiums have come and gone since the collapse of the Soviet Union. Amoco Eurasia (since merged with BP), for example, pulled out of Yamal in 1996 after several years of negotiations and feasibility studies. One reason why major Western companies are prepared to risk a return to Russia is the price of oil. Diving to \$10 a barrel in 1998, it has climbed steadily since and, despite a downward dip after the terrorist attack of September 11, 2001, has been buoyed by tensions in the Middle East.

In January 2002, Gazprom and the administration of the Yamal–Nenets Autonomous Okrug (YNAO) sent the federal government a draft law intended to make Yamal a “region of strategic interests.” The aim is to ensure the company's goal of increasing its annual output to 530 billion cm by 2010 as production from its massive Urengoi and Yamburg fields begins to decline. In doing so, Gazprom demonstrated its intention to concentrate on developing its Yamal reserves over the natural gas potential of eastern Siberia. One reason is the relative ease of transporting gas from the Yamal Peninsula. Whereas eastern Siberia is almost devoid of infrastructure, the established Yamal–Europe pipeline is capable of pumping 18 billion cm of gas annually, which could increase to 60 billion cm once a second line is completed, most likely via Poland.

There are 25 known large fields of natural gas on Yamal, with estimated total reserves of 10.4–13 trillion cm and hypothetical reserves of up to 50 trillion cm. Developing them to the extent that it would be possible to begin export of shipments in commercial quantities would require more than \$40 billion according to Gazprom's estimates from the early 1990s. By March 2003, Gazprom's board of directors had plans to adopt a “program for the comprehensive industrial development of the Yamal Peninsula fields.”

On Yamal Peninsula, the Bovanenkovo and Kharasavey gas condensate fields are the most thoroughly explored and best prepared for production. Their

development scenario calls for the construction of a liquefied natural gas (LNG) plant at the Kharasavey field and with LNG shipments by the NSR. This program will become the foundation of the company's strategic development until 2010. As such, the Yamal Peninsula and the estuaries of the Ob and Taz Rivers are expected to be major regions for the expansion of Gazprom's gas production. Adopting the Yamal Development Program will mean that other large fields, especially the gigantic Shtokmanovskoye field in the Barents Sea, will be developed in the more distant future. The simultaneous implementation of two major projects would be too costly for Gazprom and might result in intense competition between gas sellers in domestic and foreign markets.

2.4.3 *Shtokmanovskoye*

Discovered in Soviet times, the Shtokmanovskoye gas field in the Barents Sea is estimated to hold 112 trillion ft³ (3.2 trillion cm) of natural gas and 170 million bbl of crude oil and is due onstream in 2010 under a \$19- to \$21-billion development scheme. It is believed to contain the world's largest natural gas deposits. Houston-based ConocoPhillips, the world's sixth largest energy company, and TotalFinalElf each want a 25% stake in the Shtokmanovskoye project. However, as with the terminal facility in Murmansk, there is substantial skepticism among the major foreign players as to whether the project can really be developed.

As a long-term strategy, Gazprom must either focus on the development of mid-sized fields and deeper horizons of currently developed fields in the Nadym–Pur Taz District of the YNAO or undertake the full-scale development of fields on the Yamal Peninsula. The third option is the development of arctic offshore fields such as Shtokmanovskoye, but as of summer 2003 the Yamal Peninsula appeared to be the main development target until at least 2010.

3. TYPES OF IMPACT

The causes of environmental degradation commonly associated with the petroleum industry are many and have historically included drilling pads and roads and the excavation of the gravel and sand quarries necessary for their construction; blowing sand and dust from quarries and especially roads; housing; processing facilities; pipelines and power transmission lines; waste fluids from drill holes; oil spills; marine, air, and ground traffic, including rutting from off-road tracked vehicles; and seismic survey

trails. Some of these impacts (e.g., vehicle tracks) have become relatively rare occurrences in the North American Arctic since the institution of strictly enforced mitigation regulations in the 1970s. On the other hand, most of these impacts continue to be common and widespread in the Russian Arctic despite the existence of relevant regulations since the 1980s, primarily due to a lack of enforcement.

In March 2003, a panel of experts convened by the U.S. National Academies published a major review titled "Cumulative Environmental Effects of Oil and Gas Activities on Alaska's North Slope." They reported on the overall influence of 35 years of oil exploration and production activities in and around Prudhoe Bay and classified the impacts into four categories: effects on the physical environment, effects on vegetation, effects on animals, and effects on the human environment. The panel's aim was to assess not only known impacts but also future cumulative effects should production expand, for example, eastward into ANWR. The 1.5-million-acre coastal plain of ANWR occupies 5% of Alaska's North Slope and is somewhat smaller than the state of South Carolina. It encompasses six ecological zones and is considered to contain the greatest biological diversity of any protected area in the Arctic. This is probably due to its location within what is known as Beringia, a region at the crossroads of two continents during the ice ages of the Pleistocene. As a result of this geography, Alaska is home to a number of species associated primarily with North America or Eurasia as well as a fair share of endemic species.

Less tangible than the obvious impacts described previously, but also acknowledged by the panel, was that the "wilderness" or "wildlands" value of ANWR would invariably be degraded by petroleum development. This would violate, in their words, "perhaps the most prominent of the values ANWR was established to protect." Although many Alaska Natives disagree over the appropriateness of wilderness designation, there is general agreement that industrial "structures and activities [would] also violate the spirit of the land, a value that is reported by some Alaska Natives to be central to their culture."

In terms of physical and vegetation impacts throughout the Arctic, perhaps the most important factor distinguishing energy development there from that of other regions is the presence of permafrost. The late arctic-alpine biologist Dwight Billings referred to permafrost as the "glue" that binds arctic tundra ecosystems together. Permafrost exerts major influence over surface and subsurface hydrology, and thus

affects substrate temperatures and the pattern and productivity of vegetation. A large share of the global carbon budget is stored in permafrost, and there is a strong linkage between carbon dioxide (CO₂) exchange and methane (CH₄) formation and emission rates in tundra ecosystems. However, as frozen ground warms sequestration processes are affected and tundra soils can turn from carbon and methane sinks to sources. The atmospheric input of methane from permafrost regions has been estimated to be 20–40 Tg CH₄/year, which is almost 25% of the total emissions from natural sources globally. It is well-known that the types of surface disturbance associated with energy development lead to soil warming and can easily lead to thermokarst, thus threatening the maintenance of permafrost in affected areas.

In addition to direct disturbances of the ground surface, other less visible impacts can accumulate over time. These may occur independently of each other or may be exacerbated through synergy among various proximal effects. These indirect impacts, which fall under the umbrella of cumulative impacts, are well documented, and although they were formerly unforeseen, we can now predict them in many cases. According to Skip Walker of the Institute of Arctic Biology at the University of Alaska Fairbanks (UAF), cumulative impacts differ from simple additive effects in five major ways: (i) The disturbances are so close in time that the system cannot recover (time-crowded perturbations); (ii) the disturbances are so closely spaced that the effects are not fully dissipated between them (space-crowded perturbations); (iii) combinations of disturbances produce effects that are different from those of any single disturbance (synergistic effects); (iv) the effects are delayed in time or space from the original disturbance (indirect effects); and (v) large regions are affected by many seemingly insignificant perturbations (nibbling), somewhat analogous to the process of fragmentation during deforestation.

The panel convened by the National Academies concluded that the effects of oil field infrastructure are

manifest not only at the physical footprint itself but also at distances that vary according to the environmental component affected. Effects on hydrology, vegetation, and animal populations occur at distances up to several kilometers, and cumulative effects on wildland values, especially visual ones, extend much farther, as can the effects on marine mammals of sound caused by some offshore activities. All effects attributable to the structures and the activities associated with them accumulate, and many will persist as long as the structures remain, even if industrial activity ceases.

During the construction of TAPS, more than 78,500 ha (785 km²) of land was directly disturbed, of which 31,700 ha (370 km²), or 40%, was for gravel borrow sites or quarries extracted from both surface sources and streambeds. According to agromonomist Jay McKendrick, Professor Emeritus at UAF, this extraction of gravel fill for drill pads and roads is the most extensive tundra disturbance. It has been calculated that approximately 0.04% of Alaska's Arctic Coastal Plain (or 0.1% of the area between the Colville and Canning Rivers) has been affected by gravel fill. In the Russian Arctic, by comparison, exploration and development activities at a single gas field at Bovanenkovo on the Yamal Peninsula had by 1990 resulted in the loss of 127,000 ha (1270 km²) of tundra and 0.14% of the total area on Yamal. On neighboring Taz Peninsula, the figure has reached 1.5%, or approximately 6000–7000 km². Geographer Arkady Tishkov estimates that the area of destroyed ecosystems within the tundra zone of Russia comprises 8% of its total territory.

Regarding spills of liquids, from 1974 to 1977, when TAPS was first built and operated, more than 16,000 hydrocarbon spills, totaling more than 22,200 bbl, occurred along the pipeline route. During 1985 and 1986, 952 spills were reported on the North Slope, totaling 6137 bbl. In the Russian north, 7–20% of the oil extracted each year is lost through accidents along oil pipelines. A large spill of 103,000–126,000 tons near Usinsk in 1994 was reported in the world media and generated international concern, albeit briefly. However, such events are not uncommon and occur every 2 or 3 years according to Russian geographer Gregory Vilchek. Because Murmansk does not lie in the permafrost zone, its proposed oil export terminal would therefore mainly be a threat to coastal and marine ecosystems via potential spills. On the other hand, the proposed supply pipeline across the Kola Peninsula from the fields at Timan–Pechora (via Archangelsk) would certainly affect a large swath of land.

Impacts associated with energy development are complex in that various human activities can influence ecosystems simultaneously and can have both immediate catastrophic and long-term effects, leading to the aforementioned cumulative impacts. Sometimes, in practice, it is difficult to distinguish between direct and indirect impacts and scientists may use different methods for classifying disturbances. For example, Russian authors distinguish three main classes of disturbed areas: *ochagovyi* (local), *lineinyi* (linear), and *fonovyi* (spatial). The most striking example of the first type is sites

surrounding petroleum boreholes (drill sites). Transport corridors appearing in connection with road and pipeline construction constitute linear disturbances. Large territories affected by air pollution (e.g., coal burning) are examples of spatial disturbances.

4. ECOLOGICAL IMPLICATIONS OF ENERGY DEVELOPMENT

The previous section described the major existing and planned energy developments around the circumpolar North. What are the implications of these various schemes for procuring enough energy at affordable prices? Is it possible to be efficient economically as well as to mitigate major environmental degradation? What are the true environmental costs of energy extraction along the entire route from source to consumption? What if societies and cultures other than those producing or buying the energy are adversely affected during the exploration and extraction phases to the extent that their health and traditional livelihoods are threatened? These are far from trivial questions.

The following example illustrates the scope of the process of just one contemporary energy production scheme in which the full costs are not necessarily accounted for by either the producers or the end users. During a debate on whether to build a fifth nuclear reactor in Finland, some “environmental” opponents declared that purchasing natural gas from neighboring Russia would be a cleaner and safer alternative. However, others pointed out two problems with this reasoning, which they criticized as the rationale of “out of sight, out of mind.” First, much of the gas consumed in northern and central Europe is derived from fields in western Siberia, where indigenous groups dependent on traditional livelihoods (e.g., reindeer herding, hunting, gathering, and fishing) have been powerless to stop widespread habitat degradation within their ancestral homelands. Second, Russian pipeline construction standards have traditionally been well below Western levels, so that much of the gas ends up leaking into the atmosphere during transport. When burned properly, natural gas emits approximately 60% less carbon dioxide to the atmosphere than coal and approximately 30% less than oil. However, gas released directly to the atmosphere as methane has many more times the heat-trapping capacity than CO₂, making it a potentially powerful contributor to the greenhouse effect. These are just some of the

issues that need to be considered during the ongoing development of international energy networks.

Anthropogenic activities in the Alaskan Arctic up to the present time have generally created small but locally intense terrestrial disturbances ($<1\text{ km}^2$ and often $<100\text{ m}^2$), and in total they occupy small proportions ($<10\%$) of oil and gas fields. These small-scale disturbances may affect wildlife out of proportion to their spatial extent by creating micro-scale heterogeneity with patches that can either attract or repel animals. As mentioned previously, in portions of Arctic Russia, such as northwest Siberia, the realized and potential impacts are greater by an order of magnitude as a result of little or no adherence to or enforcement of regulations aimed at mitigation.

Much of the debate concerning the ecological implications of energy development revolves around habitat loss in the North American Arctic. Habitat is the place where an organism is found, including both biotic and abiotic factors. In general, the distribution of a population within its habitat at a given time is a measure of habitat value. Loss of certain portions of the species' habitat may affect the species more immediately and more severely than loss of other portions of its habitat. Loss of high-value habitats could have immediate adverse consequences for the populations that they support. Russia faces not only major losses of habitat but also the direct losses of wildlife and local species extinctions because rampant poaching of fish, wildlife, and semidomestic reindeer in and around arctic petroleum fields and along transport corridors is a major problem.

Petroleum deposits in western Siberia were first identified and tapped during Soviet times, mainly in the more accessible subarctic zones south of the arctic tree line. Those original oil and gas fields were developed recklessly with little or no concern for the technical longevity of the fields or the environmental damage resulting from the exploration and exploitation phases. It is certain that Russia will need foreign expertise in both extraction and environmental mitigation, in addition to capital investments, to develop new fields in the Arctic.

In terms of raising this capital, it makes little difference who becomes Gazprom's partner—Russian oil companies, major gas traders such as Germany's Ruhrgas, or multinationals such as BP or Shell. However, the environmental implications are potentially profound because the traditionally state-owned Russian monopolies have had few incentives for environmental protection. Western companies, on the other hand, are highly competitive and increasingly sensitive to their image. For

example, BP Amoco PLC is one of two dominant oil companies on Alaska's North Slope and has been trying to enhance its image globally during the past several years, beginning with a new sunburst logo and slogan "Beyond Petroleum" in summer 2000. Although BP pumps more than 300,000 bpd from Alaska, it has made efforts elsewhere to increase its production of cleaner-burning natural gas over crude oil and also invested heavily in solar power. Nevertheless, in December 2002, a U.S. judge tightened controls over BP by requiring the company to provide unrestricted access to its Alaskan operations to verify compliance with federal, state, and local environmental, health, and safety laws. The merger of BP with TNK raises hope for mitigation as its new Russian fields are developed.

In North America, proponents of the future development of ANWR argue that the lessons learned during the past 30 years, combined with new exploration and drilling technologies, allow engineers to design oil fields with an ecological footprint half that used in the Prudhoe Bay oil field. However, opponents point out that the switch from gravel to winter ice roads is leading to the depletion of surface and groundwater in some areas. Furthermore, the cumulative impacts from past development described previously are not yet fully understood, although it is clear that even small disturbed patches of vegetation and soils can persist indefinitely.

It is known that environmental regulation can make a positive difference. Studies of disturbance associated with exploratory well sites in northern Yukon Territory found significantly less terrain damage and thermokarst topography after the introduction of territorial land-use regulations in the early 1970s. Improvements in industry operating conditions (e.g., the use of vehicles equipped with low-pressure tires) and strict application of the land-use regulations (e.g., the restriction of cross-tundra vehicle movement to winter months) are important factors in the reduced impact observed. In this light, the potential for the most serious terrain and environmental damage is now associated with the drilling operation, which may sometimes extend into the critical summer months, and the disposal of waste drilling fluids. That the disturbances in the Yukon were neither as obvious nor as severe as in the Mackenzie River delta and the islands of the High Arctic was attributed, respectively, to the relative absence of ice-rich permafrost and the ameliorated subarctic climate conditions.

With regard to wildlife, the impacts of the oil development at Prudhoe Bay on the caribou of

the Central Arctic herd (CAH) are heatedly debated. On the one hand, the CAH has increased from 3000 animals in 1973 to 29,000 today, and the animals can be seen grazing among the oil field infrastructure in summer, even using the roads for relief from insect harassment. On the other hand, the population growth coincides with a period of favorable climate conditions that may not last. Studies have shown that mothers pregnant with calves tend to stay at least 4 km from roads and that the calving grounds have shifted away from the infrastructure over the years from the most productive habitats to less productive areas. At the same time, the Porcupine caribou herd (PCH), which migrates into ANWR from Canada, has been declining since 1989. There are concerns that if the refuge is developed the additional stress will cause an increase in calf mortality and a dramatic reduction in herd size. Concerning the proposed pipeline route, some analysts believe that using the TAPS and Alaska Highway corridor would have less impact because it would primarily use existing corridors. Opponents believe that a submerged Mackenzie Valley pipeline could endanger both marine and terrestrial ecosystems.

Concerning offshore development, the World Wildlife Fund and others have acknowledged that safe transportation with minimum pollution poses serious challenges for environmental organizations in general and marine environmental science in particular. The Arctic Operational Platform project, funded by the European Union during 2002–2004, is intended to demonstrate the development capacity of the NSR. The port at Varandei in NAO remains the only Russian Arctic stationary terminal for oil transportation and is expected to grow. Other terminals, such as the one on nearby Kolguev Island, are functioning only on a temporary basis.

There is concern that Gazprom is launching into the gas liquefaction business in an area with some of the most extreme natural and climatic conditions in the world. Frozen seas would complicate shipment of liquids and LNG via the NSR. The cold climate means that the biodegradation of oil spills on both land and in ice-covered waters would occur at a much slower rate than in more temperate ecosystems. Most of the spills of liquids on Alaska's North Slope have consisted of refined petroleum products and occurred on water or in gravel pads, although some have occurred on terrestrial vegetation. Oil does not penetrate deeply into saturated soils, but spills on dry sites are absorbed by mosses and the underlying organic material and mineral soils. Sedges

and willows are the first vascular plants to reappear following a terrestrial oil spill. Recovery from diesel spills is extremely slow, showing little change after approximately 30 years and probably after even half a century. By comparison, recovery from spills of crankcase and crude oil in the low Arctic generally occurs more quickly (<30 years) in wetland habitats. Recovery can take much longer in the colder, drier habitats of the high Arctic, not to mention in ice-covered waters.

NSR transport projects supporting the development of the Bovanenkovo and Kharasavey gas fields on Yamal require the construction of a port at Kharasavey for the import of extraction equipment and the export of liquids. At the same time, a railroad is being laid from the Ob River gulf to Bovanenkovo and a parallel service road is planned to continue from Bovanenkovo to Kharasavey. There has been a great deal of damage along this corridor, especially due to quarrying. All quarries—sand, gravel, and rubble—are slow and difficult to revegetate either naturally or with assistance and are therefore generally kept to a minimum by companies active in the North American Arctic. In the best case scenario, a complete plant cover is possible over a period of 20–30 years, although costly subsidies of nutrients, organic matter, and water are typically required to achieve this. Furthermore, the majority of plants that are either planted intentionally or eventually colonize the site naturally are not likely to be those that occurred there originally.

Another factor to consider when dealing with tundra vegetation is climate change. A warming arctic climate may offer both opportunities for revegetation efforts and risks from migrating exotic species that may compete with native species. However, with or without climatic change, we cannot expect the distribution of weedy plants, or even whole plant communities, to remain static at their northern limits. There is concern that expansion of corridors and increasing human relocation from southern regions will likely continue to provide a ready source of introduced species. A recent effort by the United Nations Environment Program to model anthropogenic impact, GLOBIO, predicted that by 2050 more than 50% of the Arctic may be affected by critical levels of disturbance. Much of the impact is expected to occur along the growing network of transport corridors and related infrastructure (e.g., housing, power lines, and airports).

Soon after the strike at Prudhoe Bay, oil companies and TAPS arranged for environmental studies of their own and supported research at the Institute of

Arctic Biology at UAF into revegetation and rehabilitation methods for tundra disturbed during the course of engineering operations and infrastructure development. Based on early experiments, seeding of three grasses [*Arctagrostis latifolia* (Var. Alyeska), *Festuca rubra* (Var. Arctared), and *Poa glauca* (Var. Tundra)] became a standard practice. On medium-texture, mesic soils, seeding and fertilizing usually produced grass stands within three growing seasons. In the long term, seeding these commercial grasses has frequently delayed natural recolonization where seeded grasses established and persisted. Examples are overburden stockpiles at gravel mines. Where such stockpiles were seeded, the commercial grasses have resisted invasion by other tundra plant species for at least 15 years. It is believed that competition from the seeded grasses is responsible for the delay of indigenous plant invasion during the process of succession.

Although there are still important gaps in our understanding of natural tundra ecosystems, it is becoming increasingly clear that there are varied and critical functions to be served by restoration procedures in disturbed ecosystems. Among the most important applications are the prevention of erosion, rehabilitation of degraded wildlife habitat, and protection of permafrost and associated benefits from carbon sequestration. Most applications to date have taken place in subarctic and low Arctic ecosystems. In the early years, research was necessarily concentrated on assisting recovery (i.e., establishing a simple plant cover as quickly as possible). Recently, however, ecological restoration has come to be seen as an aim in itself. Although plants are selected primarily to prevent erosion, many of the chosen species also provide important food for wildlife. Since it is standard practice to apply organic matter and/or chemical fertilizers to subsidize the initial stages of growth, managers need to be wary about the access of herbivores (e.g., caribou and reindeer) to sites either naturally recovering from disturbance or actively revegetated.

5. SOCIOCULTURAL IMPLICATIONS OF ENERGY DEVELOPMENT

In the discussion of the impact of development of ANWR, referred to by some as the Serengeti of the North, much of the focus has been on the area's wildlife and wildlife habitat. As a result, the human dimension is often overlooked. However, in North America, as in Russia, many communities depend to

a greater or lesser extent on fish and game for nutritional and cultural purposes. Just as onshore drilling can negatively affect calving caribou and therefore hunting opportunities, offshore oil developments may prove destructive to populations of fish, marine mammals, or birds on which local residents depend for subsistence needs. Similarly, oil or gas operations may generate forms of water or air pollution that are physically harmful to local residents. Political scientist Oran Young of the University of California at Santa Barbara observes that large-scale development involving oil or hydroelectric power in the far north may also severely disturb or distort important social institutions or central elements of the preexisting indigenous way of life, such as subsistence economies, cultural norms of sharing, and traditional authority structures. Similarly, construction projects associated with the development of northern resources may produce severe stresses on delivery systems for health, education, and welfare even in larger modern settlements.

Relative to their North American counterparts, Russia's arctic indigenous groups, the so-called small-numbered peoples of the North, lack significant political clout. Researcher Gail Osherenko of the University of California at Santa Barbara suggests that title to land may be essential to their cultural survival in the face of extensive arctic resource development. Unfortunately, despite the advent of glasnost and perestroika, there are still risks for speaking out on such issues. Some of the more vocal political activists in western Siberia have been jailed.

For groups that still practice traditional livelihoods, such as nomadic reindeer herding, the last generation has been one of adaptation to significant upheaval. The Yamal Nenets, for example, have survived first Tsarist and later Soviet dreams of establishing state and religious authority over even the most remote human populations. However, nothing has challenged them like the ongoing search for petroleum beneath their ancestral lands. The data discussed previously concerning the amount of destroyed tundra and forest-tundra are of critical sociocultural significance because virtually all of the lands once comprised reindeer grazing pasture. A culture that has changed little in 1000 years is suddenly endangered because of the havoc wrought by petroleum development. In addition to environmental degradation, the infusion of southern workers to the oil and gas fields has ushered in a host of health and demographic problems that did not exist 20 years ago.

Concerning development of the NSR, Alona Yefimenko of the Arctic Council's Indigenous Peoples' Secretariat is skeptical about measures to protect indigenous livelihoods. She believes that "practically, most of us know that real consultation and equal participation and representation by indigenous peoples requires a legal and political infrastructure that does not exist." In her view, the existing examples of energy development in Russia's north show "a very sad picture": ruined reindeer pastures, deserted oil and gas derricks, roads dissecting reindeer migration routes, etc. Nenets leader Dmitri Khorolya, president of the Reindeer Herders' Union of Russia and director of Yarsalinski sovkhoz, the largest collective management unit on Yamal Peninsula, has voiced similar alarm. In June 2001, he observed that

the vulnerable ethnic-economical systems of reindeer peoples are frequently exposed to hard market conditions, particularly where oil and gas mining has become the principal factor in the development of Arctic areas. Industrial activity in the North has resulted in the destruction of many thousands of hectares of reindeer pasture. The process is continuing. In some regions pasture degradation threatens preservation of reindeer husbandry and the anxiety of reindeer herders for their future should be heard by the world community

The pasture degradation results not only from the direct impacts by industry. Additional losses occur as the animals are forced onto progressively smaller parcels of land each year, resulting in sometimes severe overgrazing and desertification of the remaining pastures.

As in Russia, the proven and potential natural gas reserves of arctic Canada occur mainly on lands inhabited by indigenous peoples or First Nations. The difference is that the political clout of the Canadian First Nations was sufficient to halt development of the Mackenzie Valley pipeline in the 1970s. However, most of the First Nations groups that once opposed drilling on legal, environmental, and cultural grounds have abandoned their protests and now favor exploiting their region's energy resources. As a result, the government of the Northwest Territories is currently pressing Ottawa for an ambitious road and bridge construction program that would lower production costs. Stephen Kakfwi, premier of the Northwest Territories, predicts that the pipeline will soon be built and that "many communities will become twice the size, and all will have paved roads." His recently elected counterpart in the Yukon, premier Dennis Fentie, is similarly supportive of the development plan.

Native attitudes toward the development of ANWR are somewhat less homogeneous. The inland Gwich'in settlements of northern Yukon and Arctic Village in neighboring Alaska depend on the caribou of PCH and are bitterly opposed to any further development of the coastal plain. At the same time, the Iñupiat of Kaktovik on the coast of Alaska are just as strongly in favor of development.

SEE ALSO THE FOLLOWING ARTICLES

Arid Environments, Impacts of Energy Development in • Climate Protection and Energy Policy • Crude Oil Spills, Environmental Impact of • Energy Development on Public Land in the United States • Environmental Injustices of Energy Facilities • Gas Hydrates • Natural Gas Resources, Global Distribution of • Oil and Natural Gas Liquids: Global Magnitude and Distribution • Oil and Natural Gas: Offshore Operations • Oil Pipelines • Wetlands: Impacts of Energy Development in the Mississippi Delta

Further Reading

- Crawford, R. M. M. (Ed.) (1997). "Disturbance and Recovery in Arctic Lands: An Ecological Perspective." Kluwer, Dordrecht, The Netherlands.
- Forbes, B. C. (1999). Reindeer herding and petroleum development on poluostrov Yamal: Sustainable or mutually incompatible uses? *Polar Record* 35, 317–322.
- Forbes, B. C., McKendrick, J. D. (2002). Polar tundra. In: Davy, A. J., Perrow, M. (Eds), *Handbook of Ecological Restoration*, Cambridge Univ. Press, Cambridge, UK. pp. 355–375.
- Jorgenson, M. T., and Joyce, M. R. (1994). Six strategies for rehabilitating land disturbed by oil development in arctic Alaska. *Arctic* 47, 374–390.
- Khitun, O., Rebristaya, O. (2002). Anthropogenic impacts on habitat structure and species richness in the West Siberian Arctic. In "Wilderness in the Circumpolar North" (A. E. Watson, L. Alessa, and J. Sproull, Eds.), pp. 85–95. U.S. Department of Agriculture/Forest Service, Rocky Mountain Research Station, Ogden, UT.
- Nellemann, C., et al. (2001). "GLOBIO Global Methodology for Mapping Human Impacts on the Biosphere: The Arctic 2050 Scenario and Global Application." United Nations Environment Program, Nairobi, Kenya.
- Orians, G., et al. (2003). "Cumulative Environmental Effects of Oil and Gas Activities on Alaska's North Slope." National Academies Press, Washington, DC.
- Rasmussen, R. O., and Koroleva, N. (Eds.) (2004). "Assessing Social and Environmental Impacts in the North: Methodologies in Evaluation of Socio-economic and Environmental Consequences of Mining and Energy Production in the Arctic and Sub-Arctic." Kluwer, Dordrecht, The Netherlands.

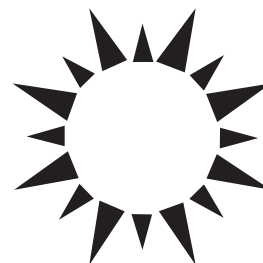
- Reynolds, J. F., and Tenhunen, J. D. (Eds.) (1996). "Landscape Function and Disturbance in Arctic Tundra." Springer-Verlag, Berlin.
- Sumina, O. I. (Ed.) (2000). "Research on Anthropogenic Impacts in the Russian Arctic," Arctic Centre Reports 29. Univ. of Lapland, Rovaniemi, Finland.
- Truett, J. C., Johnson, S. R. (Eds.) (2000). "The Natural History of an Arctic Oil Field: Development and the Biota." Academic Press, San Diego.
- Vilchek, G. E., and Bykova, O. Y. (1992). The origin of regional ecological problems within the northern Tyumen Oblast, Russia. *Arctic Alpine Res.* 24, 99–107.
- Walker, D. A., and Walker, M. D. (1991). History and pattern of disturbance in Alaskan Arctic Terrestrial ecosystems: A hierarchical approach to analysing landscape change. *J. Appl. Ecol.* 28, 244–276.

Population Growth and Energy

GAYL D. NESS

University of Michigan

Ann Arbor, Michigan, United States



-
1. A Basic Observation
 2. Early Modern Population Growth and Energy Use
 3. 1700–1950: The First Modern Population Growth Spurt
 4. 1950–2050: The Second Modern Population Growth Spurt
 5. Modern Population Growth and Energy

Glossary

crude birth rate The number of births per 1000 population.

crude death rate The number of deaths per 1000 population.

demographic transition A transition from high to low birth and death rates.

epidemiological transition A transition from high, variable death rates to low, stable death rates.

infant mortality rate The number of infant (under 1 year of age) deaths per 1000 live births; this is considered one of the best overall measures of a society's level of health and welfare.

least developed regions The United Nations designation for the world's poorest 48 countries, mostly from sub-Saharan Africa, South Asia, Laos and Myanmar in Southeast Asia, and Haiti in the Caribbean.

less developed regions The United Nations designation of poorer countries; includes most of Asia, except Japan; Africa, except South Africa; and Latin America.

more developed regions The United Nations designation for the world's wealthy, industrial nations.

The basic outlook of a long-term positive relationship between population growth and energy use is complicated by observations of different historical periods. The early modern period, in the 15th and 16th centuries, saw a new technology that harnessed wind for sails, bringing an increased population growth in Europe and Asia and a demographic collapse in the Americas. Modern population growth from the 18th through the 20th centuries was based

on fossil fuels and industrialization, with distinct differences between the North Atlantic community and the rest of the world. The present is marked by great increases in both population and energy use at the global level, with considerable variance at the level of individual countries.

1. A BASIC OBSERVATION

1.1 A Close Correlation

The basic picture of population growth in our era, the past 1000 years, is a familiar one. Figure 1 traces only the past 1000 years. The margins could easily be pushed back 100 or 200,000 years without altering the basic pattern. We have had thousands of years of slow growth, followed by exponential rates of very rapid growth over only the past three centuries. This basic pattern of growth, presented in Fig. 1, can be related to the pattern of growth in energy consumption. For this measure, a surrogate, carbon emissions, provides a picture of the past two centuries. Although this does not directly indicate energy use, it indicates the major source of energy for the past two centuries. Combustion releases carbon into the atmosphere in the form of carbon dioxide. First wood, then coal, and finally oil were the major sources of energy for the recent past (Fig. 2).

The correspondence of the two growth patterns seen in Figs. 1 and 2 is striking and indisputable. Together, these charts suggest that population growth produces increased energy consumption, or that increased energy consumption produces population growth. To be sure, if population growth were considered to occur with a constant technology (for example, using wood for heating and cooking), the juxtaposition of the growth curves would be wholly understandable. More people means more wood is needed for heating and cooking. But there are complicating conditions.

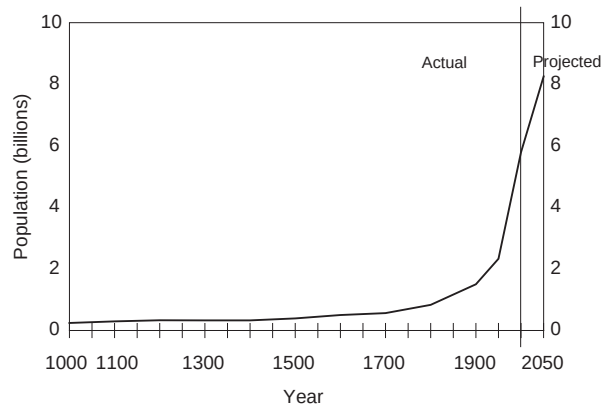


FIGURE 1 World population growth, 1000–2050. Adapted from Ness (1993).

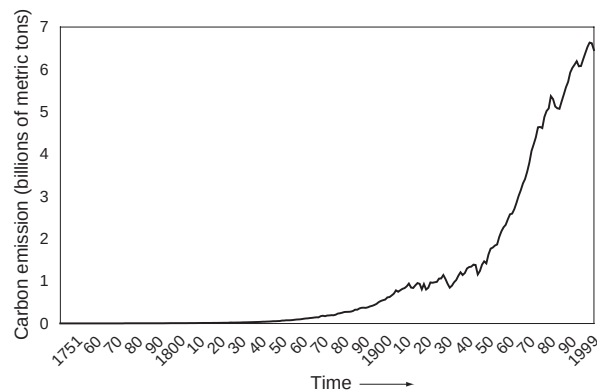


FIGURE 2 World atmospheric carbon emissions, 1751–1999. Adapted from CDIAC (2001).

1.2 Population Growth and Energy: A Complex Relationship

Ben Franklin, for one among many, showed how humans could build a more efficient stove, using less wood for more heat. Thus a change in technology can change the tight correspondence between population and energy use. This can be seen more clearly if the data in Fig. 1 are complicated by adding an indication of major changes in energy technology and social organization. This is seen in Fig. 3, which shows that the recent pattern of population growth was accompanied by a revolution in energy, from animal to sail to fossil fuels, and by an even more radical revolution in human social organization, from rural agrarian society to urban industrial society. This implies a complex change in both energy converters (animals, sailing ships, and heat engines) and sources of energy (animal feed, wind,

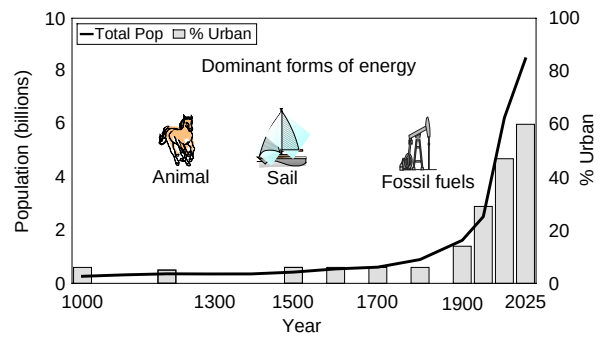


FIGURE 3 World population growth, urbanization, and energy changes.

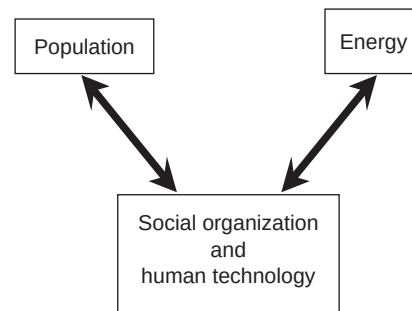


FIGURE 4 Population and energy relationships.

and carbon sequestrers). Obviously, the causal paths run in both directions. It is this observation that has led to a more complex picture of the relation between population growth and energy. It is a perspective borrowed from the discipline of human ecology. The basic model is seen in Fig. 4, but can easily be stated here: There is no direct relationship between population growth (or decline) and energy. All relationships between population and energy flow through and are mitigated by some form of human technology and human social organization.

The complex population and energy relationship can be seen in three distinctive historical periods over the past five centuries. The first period shows a small population growth spurt in the 16th century after the revolutionary progress in wind energy opened the world to a new trading system. The second shows how the modern “North Atlantic” population community grew to its present state, related to the beginnings of the fossil fuel energy system. The third focuses on the past half-century, when a new and more rapid period of population growth outside the North Atlantic Community occurred. This has been accompanied by a worldwide intensification of energy

use as the less developed countries of the world began to push for modern economic development. It has also been accompanied by both a demand and a vision for a new energy revolution that will lead to "sustained development." In all of these illustrations, it can be seen how human social organization and human technology affect the relationship between population growth and energy.

2. EARLY MODERN POPULATION GROWTH AND ENERGY USE

Perhaps one of the most dramatic illustrations of the complex relationship between population and energy is to be found in the 15th and 16th centuries. As Fig. 2 suggests, there was a great energy transformation in the 15th and 16th centuries and a slight increase in population growth rates. Until 1500, most of world trade and production were carried on by animal and human power. The Silk Route moved goods (and microbes, of which more shortly) between Asia, the Middle East, and Europe for at least the past 2000 years. In all this period, the Americas were isolated and free of Euro-Asiatic contact. That changed in the latter part of the 15th century. Columbus made the sea link between Europe and the Americas in 1492; Vasco da Gama made the sea link between Europe and Asia in 1498, and Magellan's ship completed the first circumnavigation of Earth in 1523. In the words of J. H. Parry, the west "discovered the seas" and tied the continents together in what would become a single ecosystem.

These new sea links were based on a remarkable revolution in technology, especially in naval architecture, between 1400 and 1450. This has been called the transformation from oars and swords to guns and sails. Until the 15th century, Western European, especially Mediterranean, naval architecture was based on two types of vessels. For war, there was the oar-propelled "galley," manned by slaves or freemen moving what was basically an infantry to engage in hand-to-hand combat on the sea. For trade, there were crude single-masted, square-rigged ships, often using only a large oar over the lee stern for steering. In the first half of the 15th century, the Portuguese developed the caravel, a revolutionary three-masted ship. The square sail amidships was balanced with a lateen sail in the stern and a third sail forward; all supported by rigging between the three masts. Ultimately, the hull was raised higher out of the water and fitted with canons. This new

ship proved exceptionally powerful, using wind and gunpowder to carry men and goods and to fight at sea throughout the world. It did more as well; it provided the basic design for sailing ships that grew larger and larger for the next 500 years. With this new rigging design, masts could be extended upward and increased in number, adding acres and acres of wind-catching sail. An immediate result of this new energy transformation included two dramatically different impacts on world population. The population growth rates in Europe and Asia rose from their long-term rate of less than 0.1% per annum to about 0.2% per annum. In the Americas, on the other hand, there was a massive demographic collapse.

European and Asian populations grew as new crops from the Americas quickly found their way across the seas and raised the carrying capacity of the land. Potatoes, sweet potatoes, beans, ground nuts, tomatoes, chili peppers, and corn became standard new crops in Asia and Europe, enhancing the calorie-producing power of the land, reducing mortality, and increasing population growth rates. This was an early example of what has come to be called the epidemiological transition. Historically, death rates have been both high and variable. In agrarian societies, annual changes in the weather could produce large swings in the death rate. A long winter and bad summer meant reduced nutrition and higher death rates. Better weather the following year would see an increase in nutrition and a substantial reduction in the death rate. With modernization, the death rates declined, and they also became more stable.

In return for the new crops, the Americas were given Euro-Asian microorganisms. Two millennia of Euro-Asian trade had moved diseases back and forth and had created a degree of immunity in those populations. Lacking such contact for more millennia, the Americas developed no such immunities. Measles, typhoid, and small pox proved to be deadly imports to the Amerindian populations. Within a century of Columbus' landing, the entire native population of the Caribbean was wiped out. The large and concentrated Aztec and Inca populations were literally decimated. Later, the same thing would happen to the smaller and more scattered tribal Indian societies in the rest of South and North America. This is perhaps one of the most dramatic illustrations of the complex nature of the impact of energy changes in population. The specific outcome was the result of a technological change that increased energy use, which had dramatically different impacts on different populations. The sail revolution can be argued to involve a decline in

energy use, at least in the human and animal energy used to produce and transport goods. It can also be seen as a net gain in energy: the energy gained by sails in comparison with the energy required to harness the wind. It also overcame a major limitation of the swift and maneuverable galleys, which were tied to the land because of the food and water required to feed the men who moved them. In any event, this was a technological innovation that made extensive and efficient use of a form of energy previously used only sparsely and inefficiently. This energy transformation directly produced an increase of population growth in one part of the world, and a demographic collapse in another.

3. 1700–1950: THE FIRST MODERN POPULATION GROWTH SPURT

The early modern population growth spurt brought European and Asian population annual growth rates to about 0.2% by 1700. From this point, Asian growth stagnated for the next two centuries while Europe began a new growth spurt. This was the result of what has come to be known as the demographic transition, or the transition from high to low birth and death rates. From about 1700, mortality in Europe began a long-term secular decline. Trade, agricultural, and industrial revolutions, plus a moderate global warming, gradually increased the standard of living, which lowered crude death rates from about 25–30 to 10 by about 1900. It was especially the industrial revolution that saw a dramatic rise in energy consumption. The new steam engines were powered first by wood, then the dramatic shift was made to fossil fuels, coal, and, later, oil and natural gas. Here increased energy use reduced mortality and led to increased population growth.

Although the mortality rate was decreasing throughout the 18th and much of the 19th centuries, the past levels of high fertility (crude birth rate: 30–35) remained high, producing a period of rapid population growth. In 19th-century Europe and its American extension, annual rates of population growth were about 0.8–1.0%. By the end of the 19th century, however, birth rates began to decline, coming into line with low death rates by about 1950. Thus this early demographic transition, from high to low birth and death rates, was completed in about two centuries. This coincided with the transition from rural agrarian to urban industrial society, which, among other things, transformed children

TABLE I

Regional Distribution of World Population, 1500–2050

Region	Year ^a					
	1700	1800	1900	1950	2000	2050
World population ^b	610	902	1622	2515	6057	9322
Europe	20%	20%	24%	23%	12%	6%
North America	<1%	1%	5%	7%	5%	5%
Asia	68%	69%	60%	55%	61%	58%
Africa	10%	8%	7%	9%	13%	21%
Latin America	2%	2%	4%	7%	9%	9%

^aPercentage of worldwide population in a given year.

^bMillions.

from an economic asset to an economic liability. This helps to explain the decline in fertility. In addition, the period saw the expansion of the European population to the Americas, and especially to North America, creating a new “North Atlantic” community marked by the new form of urban industrial society. Thus again, the relationship between population growth and energy appears as a complex and indirect one. Increased energy consumption came with the expansion of the machines of the new industrial society. Greater energy consumption implied lower costs of transportation and higher levels of human productivity. This facilitated urban living, implying both a decline in the economic value of and an increase in the economic costs of children. This ultimately led to lower levels of fertility and population growth. Thus, increased energy use led first to more rapid population growth, and finally to a slowing of population growth. This population growth spurt was region specific. The most rapid growth occurred in Europe and North America, changing the regional distribution of the world’s population. This can be seen in Table I.

4. 1950–2050: THE SECOND MODERN POPULATION GROWTH SPURT

The demographic transition that was completed in the Western World by the mid-20th century is now taking place throughout Asia, Latin America, and Africa. In its population reports, the United Nations refers to this part of the world as the less developed regions (LDRs). The World Bank and others refer to it as “the world of developing or less developed”

countries. The “least developed” regions, consisting basically of much of South Asia and sub-Saharan Africa, are a subgroup of the LDRs. The United Nations contrasts this with what are called the more developed regions (MDRs), i.e., the North Atlantic community, Australia, New Zealand, and Japan.

The pattern of this new demographic transition is the same as in the past: mortality decreases first while fertility remains high, producing a period of rapid population growth. Then fertility declines, coming into line with the lower mortality, and the overall growth rate slows. Although the pattern is the same, the rates, absolute numbers, and timing are radically different. First, levels of mortality and fertility in the LDRs at the beginning of the transition were much higher than they were at the beginning of the earlier MDR transition. Second, the mortality decline in the LDRs came very rapidly, following major medical advances that reduced mortality rapidly after World War II. Declines that took two centuries in the MDRs now came in two to four decades. This produced a much higher rate of population growth in the LDRs. Whereas annual population growth rates in the past transition were about 1% or less, population growth rates in this current transition are 2.5% and higher. Third, the fertility decline that is now taking place in the LDRs can occur more rapidly than it could in the past. Past fertility declines that might have required one or two generations in the past can now take place in two decades, or one generation.

The major cause of the change in the speed of the transition lies in two technological breakthroughs and associated changes in public policy and large-scale organizations. In the middle of the 20th century, a new medical technology was developed that could attack infectious diseases. Antibiotic drugs could stop infections and chemical pesticides could kill the vectors of infectious diseases such as malaria. These new techniques could be carried rapidly to the LDRs via newly created worldwide organizations such as the United Nations World Health Organization, nongovernmental organizations, and expanding private market organizations. The result was an exceedingly rapid decline in mortality and a rapid rise in population growth rates. World population growth rates stood at 1.79% in 1950/1955. They peaked at 2.04% in 1965/1970, then began to decline. For the LDRs, however, the peak in 1965/1970 was 2.51%. The MDR population growth rates peaked at 1.20% per year in 1950/1955 and have been declining steadily since then.

The modern fertility decline that is now occurring, and has been completed in many of the LDRs, can

also occur more quickly than in the past. This is due in large part to the development of non-coital-specific contraceptive technology, which emerged only in the 1960s. Like new medical technology, new contraceptive technology can be disseminated by large-scale organizations, which developed rapidly after 1960 in the form of private and public national and international family-planning programs. Moreover, these technological and organizational developments were accompanied by a radical change in national and international population policies. Most national population policies before 1952 were pronatalist. Governments supported high fertility and often banned programs and activities that promoted birth control or deliberate fertility limitation. This changed in 1952, when India became the first government to announce a national policy to reduce population growth by reducing fertility in marriage. This began what has come to be called the “antinatalist policy revolution.” Comparable policy changes became common in the rest of Asia, followed by Latin America and Africa in the next three decades. In 1968, the United Nations resolved that it is a basic human right that couples be allowed to choose the number and timing of their children and to be assisted in doing so. Today most countries in the LDRs have policies that promote or allow the promotion of fertility control programs and contraceptive distribution. Notable exceptions are found primarily in select group of Muslim countries, though many Muslim countries, such as Bangladesh, Indonesia, Iran, and Tunisia, actively promote effective national family planning programs.

Just as the first modern population growth spurt differed by world region, the current transition also differs by world region. Asia led the way in policy and reproductive behavior changes, followed by Latin America and then by Africa, which still lags considerably behind in fertility decline and thus shows continuing high growth rates. The resulting changing rates of population growth in these regions over the past 50 years can be seen in Table II.

5. MODERN POPULATION GROWTH AND ENERGY

The relation between population growth and energy currently shows two different patterns, at the global level and at individual country levels. At the macro (worldwide) level, there is a close correspondence between population growth and energy consumption,

TABLE II

World Population Growth Rates by Region, 1950–2050^a

Region ^b	Annual average (%)					
	1950–1955	1965–1970	1970–1975	1980–1985	1990–1995	2000–2005
World	1.79	2.04	1.93	1.71	1.49	1.23
MDRs	1.20	0.83	0.78	0.59	0.44	0.16
LDRs	2.06	2.51	2.35	2.06	1.76	1.48
Asia	1.93	2.42	2.25	1.85	1.58	1.26
Africa	2.17	2.59	2.61	2.87	2.54	2.33
Latin America	2.65	2.58	2.45	2.08	1.72	1.42

^a Source. United Nations, *World Population Prospects*, the 2000 Revision.^b MDRs, More developed regions; LDRs, less developed regions.

with some highly suggestive changes in the past two decades. At the micro (individual country) level, there is considerable variance in the relationship. Each requires some discussion and illustration.

5.1 Global-Level Population and Energy

For the past two centuries, there has been a very close correspondence between world population growth and energy consumption, as measured by atmospheric carbon emissions. Figure 5 shows the overall global relationship; Fig. 6 shows the relationship between carbon emissions and emissions per capita of world population. Both figures show a clear positive relationship between population growth and energy consumption. Energy consumption grew only slowly in the 19th century but showed exponential growth in the 20th century. The urban industrial society based on fossil fuels brought with it higher standards of living and substantial population growth. In the 20th century, energy consumption grew more rapidly than population did, imposing an increase in per capita energy consumption. This has begun to change in the past two decades, however, because per capita consumption, or carbon emissions, appear to have reached a plateau and may even be declining slightly. As shall be seen shortly, this declining per capita consumption has a number of causes. First, there has been a change from more polluting to cleaner fuels, so that carbon emissions can decline even as actual energy consumption rises. Second, there has been a structural change in the industrial world, showing a decline in the manufacturing and agriculture industries, based on the proportion of wealth they generate, and a rise in the service industries. The service industries are less

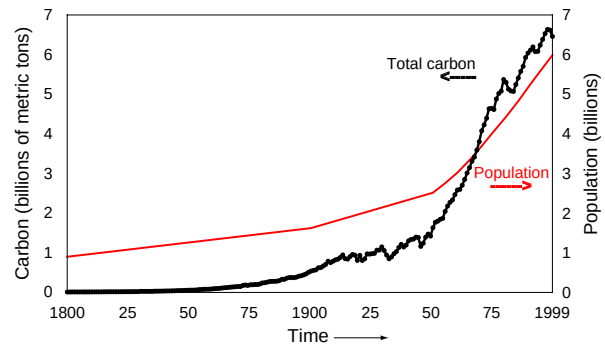


FIGURE 5 World atmospheric carbon emissions and total population.

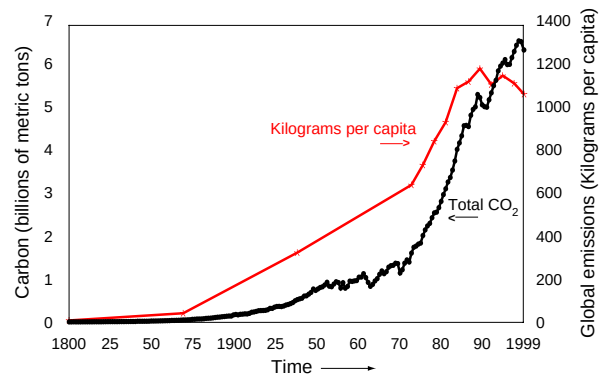


FIGURE 6 World total and per capita atmospheric carbon emissions. Adapted from CDIAC (2002) and Ness (1993).

energy demanding compared to the older manufacturing industries, thus energy use can decline even as population grows and economic development progresses. Finally, the technology of energy consumption has become much more efficient, requiring less energy for a growing population and growing wealth.

None of these changes has been universally constant, however. The specific conditions of individual countries and blocks of countries have broad effects. Individual countries differ by resource endowments, by levels of development, and by specific energy policies, both explicit and implicit. To understand these patterns requires turning to a series of individual country illustrations and comparisons.

5.2 Individual Country Variance

There are three important points to note here. First, the wealthy countries differ from the poor countries in levels of energy consumption and in the trajectories of growth. Second, wealthy countries differ greatly among themselves in energy consumption and its trajectory. Third, the less developed countries also differ among themselves in levels of energy consumption and patterns of population change.

5.2.1 *Wealthy Countries and Poor Countries*

The first point can be seen readily in Table III. The populations of the more developed regions are

growing only very slowly; their per capita energy consumption is about six times as high as that in the LDRs. At the same time, energy consumption is actually declining slightly, though it can be seen that this is due largely to Europe's substantial 20% decline from 1987 to 2000. This was primarily due to the economic collapse of the Soviet Block countries, however, as well as improvements in efficiency in other parts of Europe. Compared to the MDRs, the LDR population and per capita energy use are growing much more rapidly.

In the proportion of energy from renewable sources, there is also a substantial difference between the wealthy and poor countries. Poor countries derive about a quarter of their energy from renewable sources, and for the poorest, the proportion can reach 90%; the rich countries take only 6% from such sources. But the differences are even greater than this implies. A great part of the renewable source in the poor countries is wood, implying a low technology. In the wealthy countries, much of the renewable sources are high-technology sources such as waste recovery, hydroelectric, solar, and wind.

TABLE III

Population and Energy Consumption by World Regions^a

Region ^b	Population (millions)	Growth (%)	Total ^c	Change (%)	Per capita (kg)	Energy consumption, 1997		
						Nonrenewable (%)		
						Total	Nuclear	Renewable (%)
World	6057	1.35	9,521,506	+16	1635	86	7	14
MDRs	1191	0.3	5,827,461	-4	4505	94	10	6
LDRs	4865	1.62	3,631,617	49	803	75	1	25
Asia	3414	1.1	2,958,844	36	899	79	4	21
Japan	127	0.26	514,898	37	4085	96	16	4
Europe	727	-0.04	2,553,858	-20	3507	95	12	5
France	59	0.37	247,534	19	4233	93	42	7
MENA	432	2.28	525,927	58	1388	96	0	4
SSA ^d	651	2.56	—	—	—	—	—	—
Nigeria	114	2.74	88,652	36	853	78	0	22
S. Africa	43	1.57	107,220	15	2766	89	3	11
N. America	314	1.04	2,400,174	16	7947	94	8	6
C. America	173	1.52	1,98,317	26	1202	82	1	18
S. America	346	1.38	379,732	35	1148	72	1	22
Oceania	31	1.37	118,305	31	4038	91	0	9

^aSource: WRI *World Resources 2000-2001*, World Resources Institute, Washington, D.C.

^bMDRs, More developed regions; LDRs, less developed regions; MENA, Middle East and North America; SSA, sub-Saharan Africa.

^cThousand metric tons oil equivalent.

^dComplete data for Africa are not available; Nigeria illustrates conditions in sub-Saharan Africa.

Economic development and technology are the forces driving the changing patterns of energy consumption, and they come with low population growth rates, which have even turned negative in some cases.

5.2.2 Differences among Wealthy Countries

Most wealthy countries have seen an increase in energy efficiency, implying a slower growth of per capita energy consumption and even a reduction in the energy costs of increasing wealth. Figure 7 illustrates this process over a long time series in the United States. Over the past 130 years, U.S. real per capita output has steadily increased. At the same time, the national population has grown from 40 million to 280 million. Carbon emissions from fossil fuel emissions per \$1000 GNP have increased steadily throughout the first two decades of the 20th century then began a steady decline. Many countries experienced similar trajectories as coal was replaced with oil and natural gas and combustion processes were made cleaner and more efficient.

In the latter half of the 20th century, wealthy countries began to differ markedly in their patterns of energy consumption. As can be seen in Table III, Europe has begun a substantial reduction, 20% in the past 13 years, of total energy consumption. As noted, however, this comes primarily from the Soviet economic collapse, which has also been accompanied by a dramatic reduction in population growth rates, even to actual declines in population. Still, both Japan and Europe have much lower levels of per capita energy consumption than do the United States and Canada, for what is essentially the same standard of living. France has lowered fossil fuel consumption with a large shift to nuclear power, which now accounts for about 42% of domestic

energy consumption. Japan has also built a substantial nuclear facility, which gives it about 16% of its energy. Both Europe and Japan levy heavy taxes on gasoline consumption, leading to more extensive use of smaller and more efficient vehicles. The United States oil tax policies do not encourage the kind of efficiencies found in Japan and Europe. Thus, low population growth rates in the wealthy countries are accompanied by a wide range of different policies and technologies. These produce very different levels of energy consumption.

5.2.3 Differences among Poor Countries

There are considerable differences among poor countries in the relationship between population growth and energy. This, too, is due to different ecosystems and resource endowments, but more importantly to different policies with respect to economic and social development and environmental protection. Almost all poor countries officially espouse the aim of promoting economic development. Most also promote, either actively or passively, some form of fertility limitation to reduce population growth and promote human health. They differ remarkably, however, in their capacities to turn those aims into effective programs. Thus they differ in the population growth–energy relationship. Table IV uses three pairs of roughly comparable countries, chosen primarily for their differences in levels of welfare, as measured by the infant mortality rate. It shows their population levels and growth rates, energy consumption, and level of economic development.

Bangladesh and Pakistan are two large, poor, Muslim countries in South Asia. They share a long history of Hindu–Muslim culture and more than two centuries of British colonial rule. Both countries actively espouse aims of economic development and human welfare. Both are experiencing the recent demographic transition now sweeping the less developed regions. Bangladesh has made more progress than has Pakistan in reducing its fertility and thus its population growth rate. This has also given it a slightly higher level of human health and welfare, as indicated by its lower infant mortality rate. But its per capita rate of energy consumption is only two-thirds that of Pakistan, and its energy consumption growth rate is also only two-thirds that of Pakistan. It is slightly poorer than Pakistan, but at the moment its rate of economic development is substantially greater than Pakistan's. Here, then, slower population growth and lower economic development are

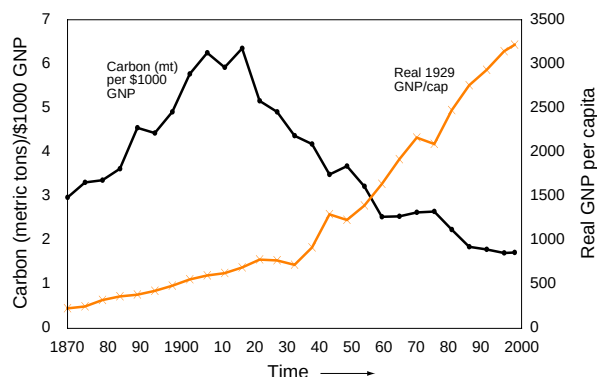


FIGURE 7 United States real gross national product (GNP) per capita and carbon emissions per \$1000 real GNP.

TABLE IV

Poor Country Comparisons on Population, Welfare, Energy, and Economic Development

Country	Population (millions)	Growth (%)	IMR ^a	Energy consumption		GNP per capita ^c	PPP per capita ^d	Growth (%)
				Total ^b	Growth (%)			
S. Asia								
Bangladesh	137,439	2.12	79	21,894	27	179	1680	3.3
Pakistan	141,256	2.66	95	42,048	39	292	1920	0.9
SE Asia								
Thailand	62,806	1.34	25	46,166	124	773	6550	0.9
Myanmar	47,749	1.48	92	12,249	14	279	600	0.5
Africa								
Zimbabwe	12,627	1.91	65	8154	9	727	2340	-9.8
Mozambique	18,292	2.31	137	6994	-1	379	1000	6.7
Latin America								
Costa Rica	226	2.48	12	1157	14	309	8080	-1.0
Guatemala	11,385	2.64	42	4433	47	421	3850	-0.06

^aIMR, Infant mortality rate.^bThousand tons oil equivalent.^cGNP, Gross national product.^dPPP, Purchasing power parity, expressing gross national income in terms of local buying power.

associated with slightly lower levels of energy consumption.

Thailand and Myanmar also share a long history of Theravada Buddhist culture in a riverine wet rice ecosystem. Myanmar came under British colonialism in the early 19th century, achieving independence in 1948, after about a century of colonial rule; Thailand managed to stave off colonial penetration through astute diplomacy, balancing British and French overseas interests. Both countries emerged in 1950 with strong aims of promoting economic development and human welfare. Thailand has been far more successful than Burma in both its development and population policies. With one of the world's most effective national family planning programs, it brought fertility from high traditional levels to below replacement in just two decades. Its associated health and education programs have given it a much higher quality of life, with an infant mortality rate about a quarter of that in Myanmar. It is wealthier and has experienced much higher rates of economic development, and its energy consumption has greatly outstripped that of Myanmar. Both countries began in 1950 with roughly equal levels of energy consumption; today Thailand's total consumption is almost four times Myanmar's and its per capita level is about three times Myanmar's. In this case, much slower population growth and more rapid develop-

ment are associated with greatly advanced energy consumption.

In southeastern Africa, Zimbabwe and Mozambique share roughly similar landscapes. Until relatively recently, Zimbabwe had a relatively progressive and effective program for economic development and social welfare. Its fertility limitation program helped bring down fertility and the rate of population growth, producing a welfare benefit, as seen in its lower infant mortality rate, which is only about half that in Mozambique. Zimbabwe is substantially wealthier and consumes almost twice as much energy per capita as does Mozambique. As in Southeast Asia, slowed population growth and greater welfare are associated with higher level rates of growth in energy consumption.

Finally, Costa Rica and Guatemala share the topical ecosystem of Central America, the long period of Spanish colonialism, and almost two centuries of political independence that failed to bring much development until the past half-century. During this time, Costa Rica embarked on a program of democratic economic development and welfare promotion that has made it one of the most advanced societies in Central America. It has had an effective fertility limitation program that has brought down the population growth rate and improved human welfare. Guatemala lagged

considerably, under both external pressures and an internally repressive government. Costa Rica is also substantially wealthier than Guatemala, yet Guatemala consumes about a third more energy per capita than Costa Rica, and its growth rate of energy consumption is about four times that of Costa Rica. Here, then, is a case in which slower population growth and higher wealth and welfare are associated with lower levels of energy consumption.

Although these are only illustrations, they show clearly how the relationship between population growth and energy consumption can be highly complex and highly varied. The near future holds promise of some striking changes in both population and energy. Population growth rates have been declining worldwide since 1965 and are expected to continue that decline. Current United Nations projections foresee a world of some 9.3 billion in 2050. The MDRs will continue to grow very slowly to 2025, when they are projected to reach 1.218 billion, after which they will decline slowly. The LDRs, on the other hand, are expected to continuing growing from their current 4.9 billion to 8.1 billion in 2050, when they are expected to be still growing at 0.57%. Because the biannual projections over the past few decades have consistently reduced the expected totals and growth rates, it is quite possible that the current estimates will turn out to be too high as well. The rising LDR population will surely increase the demand for more energy. It remains unclear what that energy will be. Certainly fossil fuels will continue to play a major part in consumption. But just as clearly, major changes can be expected. There is likely to be a gradual increase in energy efficiency, which will reduce the energy costs of economic development. There may also be a major revolution to renewable sources of energy, which will dramatically change the population–energy connection. It is expected, however that that connection will continue to be indirect, complex, and many faceted.

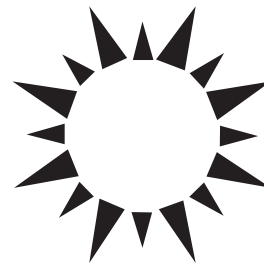
SEE ALSO THE FOLLOWING ARTICLES

City Planning and Energy Use • Consumption, Energy, and the Environment • Cultural Evolution and Energy • Development and Energy, Overview • Ecological Footprints and Energy • Economic Growth and Energy • Global Energy Use: Status and Trends • Heat Islands and Energy • Land Requirements of Energy Systems • Sustainable Development: Basic Concepts and Application to Energy • Urbanization and Energy

Further Reading

- Carbon Dioxide Information Analysis Center (CDIAC). (2001). "Fiscal Year 2001 Annual Report." CDIAC, Oak Ridge National Laboratory, Oak Ridge, Tennessee.
- Carbon Dioxide Information Analysis Center (CDIAC). (2002). "Global, Regional and National Annual CO₂ Emissions from Fossil Fuel Burning, Cement Production and Gas Flaring: 1751–1998." CDIAC, Oak Ridge National Laboratory, Oak Ridge, Tennessee.
- Cipolla, C. M. (1965). "Guns, Sails and Empire: Technological Innovation in the Early Phases of European Expansion, 1400–1700." Minerva Press, New York.
- Coale, A., and Watkins, S. C. (eds.). (1986). "The Decline of Fertility in Europe." Princeton University Press, Princeton, New Jersey.
- Livi-Bacci, M. (1992). "A Concise History of World Population [transl by Carl Ipsen]." Blackwell, Cambridge, Massachusetts.
- McEvedy, C., and Jones, R. (1978). "Atlas of World Population History." Penguin Books, New York.
- Ness, G. D. (1993). The long view: Population environment dynamics in historical perspective. In "Population Environment Dynamics: Ideas and Observations" (G. D. Ness, W. D. Drake, and S. R. Brechin, Eds.), pp. 33–56. University of Michigan Press, Ann Arbor.
- Ness, G. D., and Ando, H. (1984). "The Land is Shrinking: Population Planning in Asia." Johns Hopkins University Press, Baltimore.
- Parry, J. H. (1974). "Discovery of the Seas." Dial Press, New York.
- United Nations. (2000). "World Population Prospects, the 2000 Revision" [revised every 2 years]. United Nations, New York.
- Viola, H. J., and Margolis, C. (eds.). (1991). "Seeds of Change: Five Hundred Years Since Columbus." Smithsonian Institution Press, Washington D.C.

Potential for Energy Efficiency: Developing Nations



RANDALL SPALDING-FECHER

University of Cape Town
Rondebosch, South Africa

JOYASHREE ROY

Jadavpur University
Calcutta, India

YANJIA WANG

Tsinghua University
Beijing, China

WOLFGANG LUTZ

Energy Strategies for Sustainable Development
Ter Aar, The Netherlands

-
1. Introduction: The Context for Energy Efficiency
 2. Types of Potential for Increased End-Use Energy Efficiency
 3. Economic Potential for Energy Efficiency by Major Developing Regions

Glossary

economic potential The energy savings achieved if all new investment and retrofit were based on the most energy-efficient technologies that were cost-effective at that time.

energy intensity The amount of energy required to produce a unit of economic output.

market potential The energy saving that is expected to be realized given the existing energy prices, consumer preferences, and energy policies.

suppressed demand When the demand for energy services is low due to income, access, and infrastructure constraints, not because of consumer preferences.

technical potential The achievable energy savings that result from introducing the most energy-efficient commercial or near-commercial technology at a given time, without taking into account the costs or the life of the equipment to be replaced.

welfare potential The economic potential for energy savings when external costs and economic opportunity costs are included.

This article presents the potential for improvements in energy efficiency in developing countries. It begins by discussing the relevant development context as well as additional barriers that energy-efficiency programs face in poor countries. Then, regional overviews of the potential for energy efficiency within key sectors are presented.

1. INTRODUCTION: THE CONTEXT FOR ENERGY EFFICIENCY

1.1 The Development Context in the South: Poverty and Suppressed Demand for Energy Services

The context for end-use energy efficiency improvements in developing nations is their current energy poverty and the large share of the world's population that uses almost no commercial energy. Per capita consumption of primary energy in Africa, for example, is only 13% of the average in Organization for Economic Cooperation and Development (OECD) countries, whereas for all developing countries it is less than one-fourth (Fig. 1). The gap is even greater for electricity consumption (Fig. 2).

It is common in many poor countries for families not to be able to afford enough energy to heat their

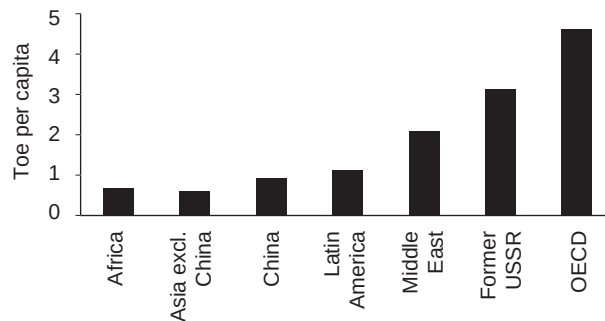


FIGURE 1 Total primary energy consumption per capita. Toe, tons of oil equivalent.

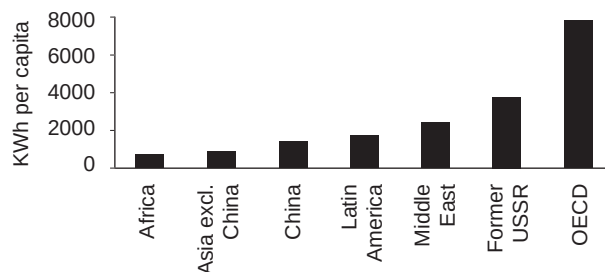


FIGURE 2 Electricity consumption per capita.

homes comfortably and cook sufficient food, let alone benefit from high-quality lighting or television. This energy poverty implies massive unmet (or suppressed) demand for energy services in these countries, with demand constrained by lack of energy infrastructure as well as income. In many African countries, for example, less than 10% of the population has access to electricity, and it is common for more than 80% of residential energy demand to be met by fuel wood.

Given these low levels of energy consumption, how do we know that there is a potential for improving end-use energy efficiency? First, it must be remembered that although developing countries' share of fossil fuel consumption is small, rapid population and economic growth will result in a substantial increase in this share in the early 21st century. From 14% of world primary energy demand in 1971, developing countries will account for 40% by 2010 if present trends continue. Even with aggressive policies to promote energy efficiency, their energy demand is likely to increase 5- to 10-fold during the next 30–40 years, resulting in a 3-fold increase in world energy demand.

Second, one measure of the potential for energy efficiency improvements in developing nations is their much higher energy intensity [energy use per unit of gross domestic product (GDP)] compared to

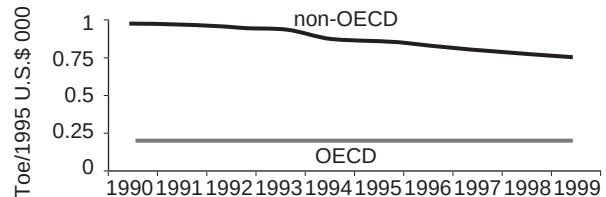


FIGURE 3 Energy intensity (energy use per unit GDP) in the developing and industrialized world.

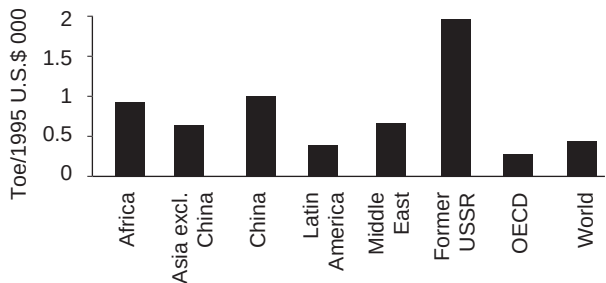


FIGURE 4 Energy intensity by region (energy use per unit GDP).

that of industrialized countries. As Fig. 3 shows, based on International Energy Agency data, energy use per dollar of GDP in the developing world is almost three times that of the OECD countries. If we consider the difference in purchasing power of developing currencies, this picture changes, with developing countries being only approximately 30% more energy intensive. For products that are traded, particularly those exported by developing countries, purchasing power parity does not really matter. In other words, if India is exporting steel and using two or three times as much energy to make a dollar's worth of steel, that is a very high cost and a competitive disadvantage. Of course, it also represents an opportunity to increase energy efficiency in a cost-effective manner. Even within the developing world, however, there are vast differences in energy intensity, as shown in Fig. 4.

1.2 Additional Barriers to Energy Efficiency in Developing Nations

Although barriers to the uptake of energy efficiency have been generally described elsewhere, they take a different form in many developing countries, and some are unique to developing countries. Before discussing different regions, however, it is useful to highlight some of the more significant barriers in developing countries.

1.2.1 Affordability and Financing

Investment in energy efficiency improvements is often constrained by the limited and irregular cash flow in poor households and difficulties in accessing additional finance; they frequently have to rely on hire purchase agreements or “loan sharks.” Poor households sometimes invest in fuels and appliances that, ironically, are inefficient in both energy and economic terms. Poverty dictates overlooking long-term factors such as the life cycle costs, efficiency, and safety of an appliance.

For fuel purchases as well, households are often constrained by income flows, with a decision on purchase depending on the availability of money at a particular time. When there is little, kerosene and coal stoves are used because one can buy these in small amounts. Householders know that buying fuels this way is expensive, but not enough money is available to buy in bulk.

Even within industry, a shortage of capital or high interest rates are significant barriers to normal equipment upgrades, not to mention the purchase of more expensive energy-efficient equipment. This means that the rate of capital stock turnover is lower than for similar industries in industrialized countries.

1.2.2 Information

Information and awareness programs need to take into account the wide range of stakeholders involved in the delivery of energy efficiency services. For example, poor consumers may have low levels of literacy or need materials in local languages. Other stakeholders include local government, utilities, manufacturers, developers, builders, and nongovernmental organizations—all with different informational requirements. Often, awareness programs tend to provide broad conceptual information when the target audience requires more sophisticated information. Where people are already aware of broad notions of energy efficiency, for example, information may be needed on practical implementation, such as how to build an energy-efficient house or install a ceiling, where to access financing, how to mobilize capital to finance energy efficiency for the poor, and how to facilitate community participation.

1.2.3 Physical Access to or Availability of Fuels

In certain circumstances, low-income households are unable to secure the best mix of fuels because certain fuels are not readily available. For example, although kerosene networks are good in many countries, with both fuels and appliances being readily available, many poor household do not have electricity

connections. Although a few developing countries have started to develop residential bottled gas networks, this is the exception rather than the rule. The coverage of distribution networks is often weak and inhibited by poor transport infrastructure and a lack of access to transport.

1.2.4 Split Incentives: Construction, Ownership, and Use

In the delivery of housing, those making the initial capital investment in constructing a house are generally separate from those who will live in it and pay the operating costs. It is not common, therefore, to find developers investing in energy efficiency. The consequence is that the low-efficiency house has high operating costs. This is a major problem in developing countries with large public housing programs because the government priority is to deliver as many houses as quickly as possible; for both government and contractors, the tendency is to cut corners and minimize initial cost. It is also a problem in industrialized countries.

1.2.5 Lack of Tenure and Urban/Rural Commitment

In many urban informal settlements, appliance ownership is constrained by space and tenure problems. Lack of space, for example, means that in informal electrified settlements and backyard shacks, more households own two-plate electric stoves than stoves with ovens. More important, the tenure problems in informal unplanned settlements and shacks play a direct role in purchasing electrical appliances or other expensive investments in efficiency. Migrant workers continue to play a large role in many countries' urban communities. They are committed to their rural households and view urban life as temporary and thus tend to save or remit money for the maintenance of the rural households, which obviously implies an even lesser willingness or ability to pay a higher cost for efficient appliances.

1.2.6 Multiple Fuel Use and Household Needs

Multiple fuel use means that households use more than one fuel for the same end use. It is common to use gas, kerosene, coal, and/or electric stoves for cooking. Gas might be used for specific tasks (such as cooking special quick foods) and kerosene appliances used for foods that take a longer time to cook. Therefore, although gas may be a more energy-efficient cooking fuel than kerosene or electricity (and healthier than kerosene), households may not completely stop using kerosene after they buy a gas

stove. For example, they may use kerosene for cooking certain foods or if they run out of gas and do not have enough cash to buy an entire gas canister since kerosene can be purchased in small quantities. Also, in some contexts one appliance serves more than one purpose, as in the case in which a kerosene or coal stove is used for cooking and space heating.

1.2.7 The Symbolic Value of Appliances

Not surprisingly, poor consumers do not simply consider the economics of appliance choices when making decisions: Symbolic value can be as important as functional value. For example, many formal households tend to replace their nonelectrical appliances with modern and sophisticated appliances immediately after electrification, not simply because electricity is cleaner and a more convenient fuel. There is a general perception that nonelectrical appliances are not appropriate for formal households. Having electric appliances attracts respect and envy, as symbols of modernity and comfort, and many people will go far to acquire them: The bigger the appliance, the better. Therefore, consumers may not be attracted to smaller, more efficient appliances unless other features enhanced the sense of “modernity.”

1.2.8 Technical and Managerial Skills

Compared to industrialized countries, the industrial and commercial sectors in developing countries have lower levels of skill in both the technoeconomic dimension of energy efficiency and the managerial skill to develop and administer programs. There may be many qualified analysts and professionals, but resources are spread too thinly across huge economies and social needs. This is particularly a problem for small to medium firms.

2. TYPES OF POTENTIAL FOR INCREASED END-USE ENERGY EFFICIENCY

As a background to describing the potential for increasing end-use energy efficiency, it is important to understand the different types of energy efficiency potential. The most important categories are technical, economic, and market (Fig. 5):

- Technical potential is the achievable energy savings that result from introducing the most energy-efficient commercial or near-commercial technology at a given time, without taking into account

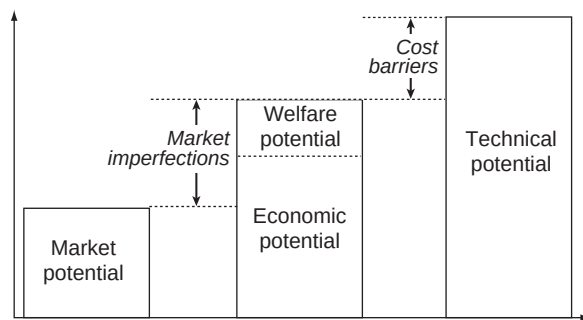


FIGURE 5 Types of potentials for energy savings.

the costs or the life of the equipment to be replaced. Realizing the full technical potential means scrapping existing appliances and equipment, although sometimes a phased technical potential is described, meaning that all new equipment is replaced with the most energy-efficient technology.

- Economic potential is the energy savings achieved if all new investment and retrofit were based on the most energy-efficient technologies that were cost-effective at that time. Cost-effectiveness at a minimum must include the energy and operating savings on the benefits side and the additional capital, any additional maintenance or control costs, and future replacement costs on the cost side. There are several perspectives on cost-effectiveness. For an individual firm or household, we would only consider the costs and benefits mentioned previously. In a macroeconomic perspective, however, we would need to consider the costs of implementing the energy efficiency programs. We would also include the benefits that could accrue to the broader economy, such as changes in technology costs over time through learning and economies of scale, or other spillover effects and nonenergy productivity gains. Finally, if we consider the societal perspective, we should include the avoided external costs of the energy saved by end-use efficiency measures. This would increase the economic potential since the benefits (avoided costs) increase. This is called the welfare potential of energy savings.

- Market potential is the energy saving that we expect to realize given the existing energy prices, technology, consumer preferences, and energy policies. This definition must take into account the social obstacles to energy efficiency and other market imperfections (i.e., no internalizing external costs) that block the realization of the full economic potential. In addition, market potential may consider whether energy users choose to use their energy

savings to buy more of what is essentially a less expensive service. This is called the take back or rebound effect.

These definitions also help us understand that the role of energy policy in promoting end-use energy efficiency is to both remove the market barriers that keep market potential below economic potential and address cost issues (e.g., through targeted research and development) that can bring the economic potential closer to the phase-in technical potential. Later, when we discuss the potential for energy-efficient improvements, we refer to economic potential unless otherwise noted.

3. ECONOMIC POTENTIAL FOR ENERGY EFFICIENCY BY MAJOR DEVELOPING REGIONS

3.1 India

3.1.1 Trends in Energy Intensity and Consumption per Capita

India, the world's second most populous nation, has seen its population explode from 300 million in 1947 to approximately 1 billion today. Energy consumption has grown rapidly through the process of development. Aggregate consumption has increased approximately 50% during the past decade. The share of commercial energy consumption increased from 26% in 1950–1951 to approximately 60% in 1994–1995. With an elasticity of commercial energy demand (i.e., the percentage change in energy demand for a 1% change in income) of approximately 0.92, and with the objective of rapid economic growth of 6% per annum, a high growth in commercial energy demand is expected. India is the second largest commercial energy consumer within non-OECD East Asia (after China), comprising 19% of the region's total primary energy consumption.

The current low per capita energy consumption (3% of that of the United States and 9% of that of Korea) is an indicator of huge potential for growth in energy demand. Despite an annual approximate 5% growth rate in electricity generation, unmet demand for electricity continues. Although 75% of India's population lives in rural areas, only 29% of rural households have electricity supply and there is no power supply network in many rural areas. In addition to geographical constraints, sparse distribution of dwellings, low ability to pay, and low load make extending the distribution network either impossible or uneconomic for many remote villages.

Noncommercial energy still meets 40% of the total energy consumption, which shows the extent of demand that awaits to be met by commercial and more efficient forms of energy.

Energy demand in India has always been greater than supply. Thus, in view of shortages and restrictions, the past trends in consumption of commercial energy do not really represent the growth of energy demand but reflect the growth of its availability. Energy consumption by fuel is dominated by coal (54%) and oil (31%). There is wide scope for enhancing end-use energy efficiency. In addition to replacing outdated technologies and removing market barriers such as pricing distortions, overall economic potential would increase by simply changing the scale of production of many production activities. Such potential exists in all of the energy-using sectors.

Sectoral demand for energy arises mainly from lighting and cooking demand in the residential sector, irrigation and other operations in the agricultural sector, transport of passengers and freight in the transport sector, and fuel requirements in the industrial sector. Fig. 6 shows the use of energy by sector, whereas Fig. 7 shows the trends in energy intensities of these sectors. Decomposition analysis of the energy intensity trends shows that in India increasing economic activity is the main driving force for rising energy consumption, not increasing energy

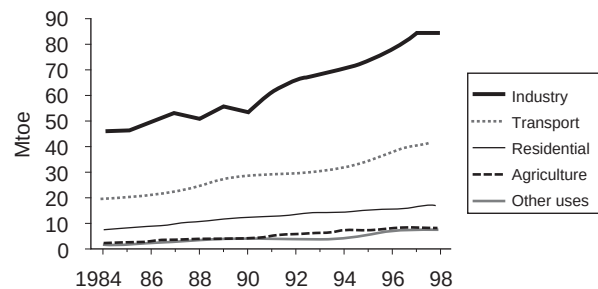


FIGURE 6 Trends in consumption of commercial energy in India, 1984–1998.

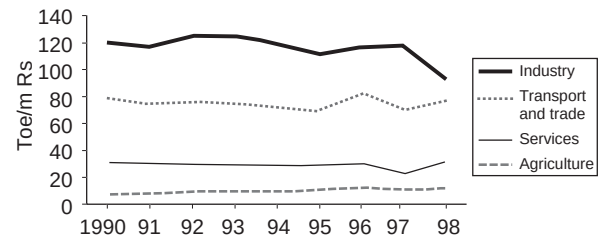


FIGURE 7 Sectoral energy intensities in India (energy/real GDP).

intensity. Since it is difficult to restrict energy demand due to increased output or activity directly at the early stage of development, improved energy efficiency is essential.

A number of efforts to improve energy efficiency, especially through exploration of technical potential, have resulted in a declining trend in energy intensity. The Eighth Plan (1992–1997) included a National Energy Efficiency Program that has the potential to save 5,000 MW in the electricity sector and 6 million tons of oil in the petroleum sector over the 5 years. India established an energy efficiency center as a major agency to implement programs and an energy conservation fund to finance them. On the other hand, electricity prices are far below marginal costs throughout most of the country, as are prices for natural gas and coal. This is a substantial barrier to energy efficiency investments and essentially subsidizes energy sources with environmental and health costs from pollution. One of the major challenges in a country with such high unmet demand is how to correct prices and provide incentives for investment in energy efficiency while protecting the social needs of the poor. A subsidy program for efficient vehicles or appliances without a move toward cost-reflective prices, for example, could substantially increase overall energy demand, as consumers buy more of a much needed service, and not reduce environmental impacts from the sector.

3.1.2 Industrial Sector Potential

In general, Indian industry is highly energy-intensive and its energy efficiency is far below that of industrialized countries. For example, in nitrogenous fertilizer manufacturing, older manufacturing units generally require 8.4–10.5 tons of oil equivalent (toe) per ton of ammonia produced, whereas the standard achieved in a plant in Bristol, England, is only 4.6 toe. An improvement in energy efficiency also includes energy conservation undertaken by a producer or a consumer that reduces losses of energy. In the electricity sector, it has been estimated that older power plants in many developing countries consume 18–44% more fuel per kilowatt hour of electricity produced than plants in industrial countries. Since India has vast reserves of coal, this is the primary fuel in the industrial sector. India has always relied heavily on coal for generating electricity with a low average conversion efficiency of 23%. There is potential for at least a 10% improvement in the conversion efficiency of existing plants. Greater energy efficiency can be implemented through technological enhancements, more research and

development, improved energy management, and better operational practices.

A World Bank-sponsored study (1995) of energy demand in five major Asian countries (China, India, Indonesia, Korea, and Thailand) concluded that the industrial sector has the largest potential for energy conservation. This sector accounts for the largest share of final energy consumption in India and is dominated by energy-intensive industries. Higher energy intensities are ascribed to (i) slow capital stock turnover and utilization of out-of-date processes, (ii) poor energy management, and (iii) lack of incentives for firms to invest in energy efficiency due to low energy prices. In addition, scale is a major issue in many industries. In nearly all states in India for the cement and paper sector, and in a few states for aluminum, increasing the scale of production could reduce inefficiency in energy use.

We can also consider a recycling strategy in which secondary materials are fed into the original process, thereby lowering per unit energy consumption. Examples include the substitution of waste paper for paper pulp and using recycled glass and plastics. To improve overall industrial energy efficiency and reduce pollution, there is greater potential from adopting available clean technologies at the beginning of industrial operation rather than “end-of-pipe” solutions, which can actually increase overall energy use.

Technical energy efficiency improvement potential is shown by comparing specific energy consumption (i.e., energy consumption per unit of output) in industrial subsectors. For example, the energy efficiency of ammonia plants can be increased by revamping and modernization. A typical energy efficiency revamp of a plant would reduce the energy consumption for individual plants installed before 1980 by 0.8–2.2 Mtoe and for plants installed between 1980 and 1990 by approximately 0.6–0.7 Mtoe, depending on the feedstock used. The age of the technology, the scale of the plant, and management practices have a major impact on energy efficiency of the overall process. Energy savings potentials are highest for naphtha and fuel oil/low-sulfur heavy stock-based plants built before 1980. With current installed capacity of 5.09 Mt production, energy savings in pre-1980 ammonia plants alone would account for 6.02 Mtoe per year.

In the iron and steel sector, there is a large potential for efficiency improvement. Worldwide, specific primary energy consumption is decreasing with rising scrap-based electric arc furnace production. In India, however, due to a scarcity of scrap metal, this may not be a good option. Still, efficiency

improvement may be achieved indirectly by improving power-generation efficiency that feeds iron and steel plants through modernization and captive combined heat and power generation. In addition, technology conversion (e.g., to induction furnaces) and retrofitting, recycling, and waste heat recovery promise major savings. Although most measures are cost-effective and provide net benefits within a short time, only a few have been or are currently being implemented in the iron and steel sector. Lack of dissemination of information on energy-efficient technology as well as specific information on saving and benefits of saving potential contribute to the low uptake of energy efficiency improvements.

For the cement industry, there is not much scope for improvement in dry processes, but there is more in semidry and wet processes. Besides the technical potential when compared to world best-practice technology, efficiency improvements can be achieved through increased use of additives such as fly-ash and blast furnace slag in manufacturing cement. Structural change toward more production of portland slag cement reduces energy consumption, and using a higher share of blast furnace slag compared to clinker leads to lower energy intensity as well. In India, 20% of final energy could be saved by this twofold structural change, which would decrease the clinker-to-cement ratio by 20%. It has also been estimated that better maintenance and monitoring of plant activity can minimize downtime of machinery and plant, thus avoiding excess energy needed for restarting the process.

In the aluminum sector, the potential for energy efficiency improvement varies across companies. For example, whereas the National Aluminum Corporation could achieve 17%, the Hindusthan Aluminum Corporation could achieve 40%. Energy savings potential in smelting range from approximately 16 to 30%. Better monitoring and control can conserve

an additional 2 or 3%. Potential from retrofitting ranges from 10 to 15%, and the costs of this retrofitting and modernization are 25–35% lower than the cost of building an entirely new plant. Table I summarizes the industrial sector potentials for energy efficiency improvements by comparing Indian specific energy consumption to worldwide best practice. Worldwide best practice reflects the best available technology adjusted for India-specific structural factors. For paper, best practice reflects the average energy consumption abroad.

Research on the responsiveness of the Indian industry to changes in energy prices indicates that price-based policies could be effective in improving energy efficiency. Pricing policy in the fertilizer sector, for example, has shown that energy price increases would push less productive and inefficient plants (mostly smaller ones) out of the market or force them to immediately improve productivity and efficiency. It has been estimated that with all other economic variables constant, India industry could achieve worldwide best practice in specific energy consumption by 2005 or 2010 with an average annual nominal energy price increase of approximately 6 or 4%, respectively, measured as an increase in the fuel price index relative to other input prices.

3.1.3 Residential and Commercial Sector Potential

In the residential and commercial sector, the demand for energy is mainly for lighting, cooking, space cooling, and, to some extent in higher altitudes, space heating. The residential sector in India accounts for 40–50% of the total energy consumption. For cooking, 77% of households in India use traditional fuels such as firewood, animal dung, and agricultural residues, 0.3% use electricity, 4% use coal or charcoal, 8% use liquefied petroleum gas (LPG), 7% use kerosene, and 0.5% use biogas. Experience

TABLE I

Specific Energy Consumption in Indian Industry and Worldwide Best Practice, 1993

Sector	Units	India		Worldwide best practice	India
		Average energy consumption	Best practice (marginal plant)		Savings potential (%)
Aluminum	GJ/ton aluminum	86–100	87	65.2–81.9	5–35
Cement	GJ/ton cement	4.00	~2.93	3.06	24
Fertilizer	GJ/ton ammonia	39.2	36.4	28	29
Iron and steel	GJ/ton crude steel	35.4	31.4	17.1	52
Pulp and paper	GJ/ton paper	51.6–80.0	NA	32.0–40.9	~43

tells us that in the ladder of fuel-use transition, households usually move from traditional fuels to kerosene/coal (depending on availability), LPG, and electricity; the use of traditional fuel is indeed an indicator of unmet demand for more efficient fuels.

In India, urban and rural energy problems differ. In urban areas, introducing energy-efficient lamps, better space-cooling devices, and solar passive building structures can enhance efficiency in electricity and fuel use enormously. In the absence of any major demand-side management through policies or appliance standards, however, there is hardly any incentive for the households or commercial sector to enhance efficiency in use. Adopting energy-efficient appliances could provide up to 50% savings in electricity consumption in these end uses. This is particularly important in India, where demand for electricity is expected to grow by 7% per year and the gap between peak demand and peak supply today is more than 11,000 MW.

In rural areas, where almost 75% of the Indian population lives, biomass (primarily firewood) accounts for more than 90% of energy needs. It may not be possible to replace this entirely with more modern energy forms and developing modern biomass technologies will be a key element for addressing the needs of rural areas. One of the important constraints in the wider utilization of renewable energy is the high initial capital cost. India has potential for renewable energy generation, particularly wind and hydropower, and programs in these areas, aided by Global Environment Facility (GEF) funding, are under way. At the same time, India presents a unique opportunity for the large-scale commercial exploitation of biomass gasification technology to meet the needs of the agricultural and rural residential and small-industries sector. The gasifier power could solve the problems of technical and economic inefficiencies to a large extent, thereby reducing the social cost of power supply. Currently, commercial activity in most rural areas depends on power supply from the privately owned, decentralized, diesel power-generating sets, whereas residential lighting is typically met with stand-alone kerosene lamps. Use of kerosene for lighting is particularly inefficient, with only 1% of the energy in the kerosene producing useful light. The Ministry of Non-Conventional Energy Sources aims to replace low-efficiency kerosene lamps in rural areas with efficient renewable lighting technologies.

3.1.4 Transport Sector Potential

Transport shows a constantly rising trend in fuel demand. Decomposition analysis shows that it is the

rising demand for overall transportation service that is driving up fuel demand. With relatively low substitution possibilities and high income elasticity for privately owned vehicles, it is expected that fuel demand in the transport sector will increase along with vehicle ownership for years to come. In recent years, fuel switching has been attempted by introducing compressed natural gas in public transport in metropolitan cities. Fossil fuel-based transportation may also be reduced through development of the rail network. Pricing reform is playing an important role in reducing inefficient use of energy. In terms of technical interventions, studies have shown that inefficient use of fossil fuels often results from poor design of public buses, which are not using best-practice technologies. Efficiency in transportation fuel can also be increased remarkably by improving the urban road conditions and reducing traffic congestion.

3.1.5 Agricultural Sector Potential

In the agriculture sector, power use for irrigation holds significant potential for improving energy efficiency. Currently, diesel generator-based water pumps with efficiencies of less than 33% are used to draw underground water; an adequate electricity supply would allow a shift to more efficient electric pumps. Where electricity is now available for agriculture, it is heavily subsidized and charges are not linked to consumption, leading to inefficient use. Correction of prices would lead to increased efficiency through better water and irrigation management. Just an important, more efficient use of water for irrigation would decrease agricultural energy demand for the same agricultural output.

Table II summarizes the potential for energy efficiency improvements in India based on a wide range of studies.

3.2 China and Southeast Asia

3.2.1 Trends in Energy Intensity and Consumption per Capita

China has demonstrated significant improvements in energy efficiency since the 1980s. Energy use per unit GDP declined from 2.18 toe per thousand U.S. dollars in 1980 to only 0.73 in 2000, as shown in Fig. 8. During this period, commercial energy consumption increased 2.3 times, whereas the GDP increased more than six times. However, commercial energy intensity in China is double that of OECD countries and triple if noncommercial energy is included (Fig. 4). Many factors affect energy intensity, including economic

TABLE II

Summary of Energy Efficiency Potentials for India

Sector	Technological options/area	Economic potential (% of current use)
Industry	Fertilizer	29
	Cement	20–24
	Electrical	17
	Thermal	27
	Pulp and paper	20–40
	Textiles	23
	Iron and steel	15–50
	Aluminum	5–35
	Refineries	8–10
	Brick-making	15–40
	Foundries	30–50
	Industrial cogeneration	3500 MW (sugar)
Residential	Lighting	10–70
	Refrigerator	25
	Air-conditioning	10
Agriculture	Pump sets	25–55
Transport	Two- and three-wheelers	25
	Cars	7.5–10
	Trains (diesel)	5–10
	Trains (electric)	5–10

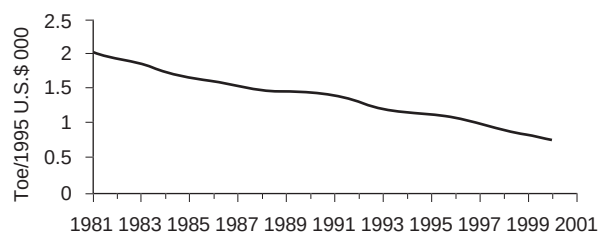


FIGURE 8 Energy intensity (energy/GDP) trends in China, 1981–1999.

structure, product mix, exchange rates, energy resource mix, and technical efficiency. In general, industrial sectors, especially heavy industry such as the iron and steel industry and cement industry, consume more energy per unit of economic output than do light manufacturing, finance, and service sectors. In China, for example, the relative energy consumption per unit of value added in primary (e.g., resource extraction and beneficiation), secondary (e.g., manufacturing and construction), and tertiary (e.g., finance and services) sectors was 12, 68, and 20, respectively, in 1999.

Changes in economic structure have contributed significantly to China's energy intensity improvement. Fig. 9 shows that economic structure has

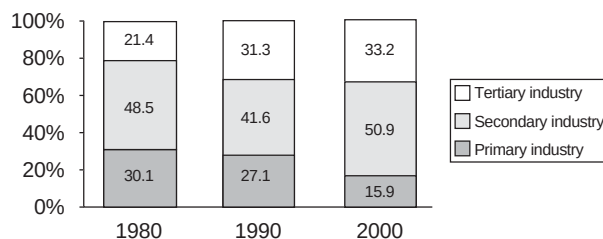


FIGURE 9 Composition of GDP in China, 1980–2000.

changed in China, with a decline in primary industry replaced by rapid growth in the tertiary sector. This trend is expected to continue, and the share of tertiary activities in the total economy will increase. Even within the secondary sector, subsectors with higher value-added and lower energy intensity, such as computer manufacturing and telecommunications equipment production, are developing faster than the traditional heavier manufacturing industries.

China is one of the few countries where coal is the dominant resource for energy use, which is part of reason for China's high energy intensity. For example, even using the latest technology, the highest energy efficiency for combined heat and power generation based on coal as a resource is 60%. Natural gas, on the other hand, can obtain conversion efficiencies of more than 80%. China is currently promoting clean energy production and importing cleaner technologies, which will result in energy efficiency improvements. In Southeast Asian countries, energy intensity is much lower than that of China. Because of the financial crises in the mid-1990s, however, the primary energy consumption per GDP increased, as shown in Fig. 10.

To provide a regulatory framework for energy efficiency interventions, many Southeast Asian countries and China have enacted a range of energy efficiency laws, with differences based on culture, political ideology, and existing institutional structures. Although the scope varies, common provisions include mandatory auditing for industries, mandatory energy building codes, developing performance standards for energy-using equipment, information dissemination, and capacity building. In most cases, implementing the legislation is carried out by existing energy agencies involved in energy efficiency promotion. In Thailand, an energy conservation fund to finance activities and programs for energy efficiency promotion has been set up, whereas for other countries the law specifies budget allocations for energy efficiency programs. For countries without specific legislation, energy efficiency activities are limited to programs that do not require specific

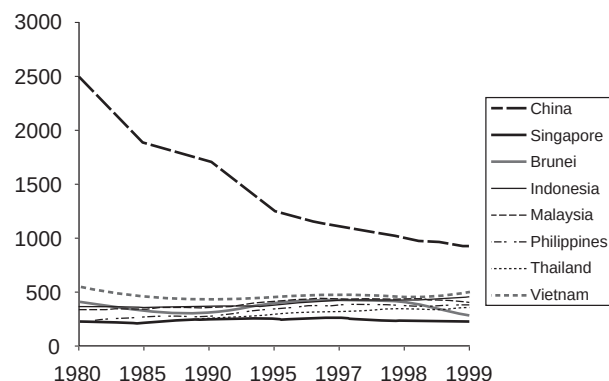


FIGURE 10 Primary energy consumption per unit of GDP.

regulation, such as publicity campaigns, demonstration projects, and programs that can be funded by external grants and loans, government budgets, or payments by consumers and end users. Some energy efficiency programs, such as mandatory appliance standards, appliance labeling, and building energy codes in the Philippines, Malaysia, and Singapore, are implemented through regulation but without reference to any specific law. A wide variety of policy instruments are used in promoting energy efficiency and accelerating the diffusion of energy-efficient technologies. These include appliances standards and building energy codes, appliance labeling, and financial incentives.

Since opening up to the outside world, the Chinese government has paid significant attention to energy conservation projects, including energy policy objectives to “carry out conservation and development of resources simultaneously but giving priority to conservation.” To address energy shortages, environmental protection, and sustainable development, the Energy Conservation Law was passed in 1997, providing broad guidance for energy efficiency improvement in all sectors. Since then, governments at national, provincial, and local levels and industrial associations have issued a number of energy efficiency policies and adopted measures for the law’s implementation.

3.2.2 Industrial Sector Potential

China’s industrial sector is extremely energy intensive and accounted for 76% of the country’s total energy used in 1997. It is estimated that the overall final energy use efficiency in the industrial sector was 46% in China, 20% lower than that of OECD countries. The largest energy-consuming industries are chemicals, ferrous metals, and building materials. The energy use per unit of product output provides a

better comparison than energy use per unit of production value added. Because of outdated production technologies and processes, slow capital stock turnover, small-scale operation, and poor management and training of staff, most industrial enterprises in China consume more energy to produce a unit of product than applies in developed countries. Even within China, the product energy intensities are quite different in various enterprises. Table III compares energy intensity for selected products between developed countries and China.

There are several technologies available for energy efficiency improvement in industrial sectors. For example, in the cement industry, wet to dry process conversion, on-site cogeneration of heat and power, and coal ash utilization can significantly decrease energy consumption. Based on the available research, it is reasonable that energy efficiency in industry could be increased by 15–20% in China, with a similar situation in Southeast Asia (Table IV).

3.2.3 Residential and Commercial Sector Potential

In 1999, the Chinese residential sector consumed 78 Mtoe commercial energy, accounting for 14% of total final consumption of commercial energy. Including biomass, total residential final consumption was 289 Mtoe. The commercial sector accounts for a further 3% of commercial final consumption. Most residential energy consumption is for space heating, followed by cooking and lighting. Electricity consumption for appliances is increasing rapidly, and the use of air-conditioning is especially important for commercial buildings and it is being used increasingly in households as well.

The rate of construction of new buildings is high due to increases in population, rising standards of living and demands for more living space per person, and the continuing process of urbanization as people migrate toward areas with greater economic development. The majority of the new buildings, however, have poor energy performance. This is not unlike what happened in Europe in the years of reconstruction after World War II. The urgency of supplying new housing and commercial space, pressure to hold down initial investment costs, and low energy prices mean that little attention is paid energy performance. Designs are not adapted to local climatic conditions and ignore the potential for day lighting in commercial buildings.

As a result, energy costs are much higher than expected on economic grounds, even without considering the positive environmental and health impacts of better building design. In the northern part

TABLE III

Energy Intensity in China and Industrialized Countries for Selected Products

Product	Specific energy use	Country	Year
Steel (GJ/ton steel)	60	China (average)	1980
	41		1996
	32		1999
	21	China (best)	2001
	25	United States	Mid-1990s
	18	Japan	
Cement (GJ/ton cement)	21	Sweden	
	6	China (large and medium-sized)	1980
	5		1996
	4	United States	Mid-1990s
Paper (GJ/ton paper)	3	Japan	1990
	46	China	1997
	21	OECD countries	
Ethylene (GJ/ton)	36	China	
	26	OECD countries	
Glass plate (GJ/box weight)	0.8	China	
	0.4	OECD countries	
Ammonia (GJ/ton)	41	China	
	28	OECD countries	
Copper (GJ/ton)	40	China	
	24	OECD countries	
Cotton yarn (kWh/ton)	2300	China	
	2100	OECD countries	

of China, energy consumption for space heating during winter is approximately twice as high as in areas with similar climatic conditions in industrialized countries. As the new building energy code is implemented, new buildings are expected to have a 20% energy use savings compared to existing buildings.

3.2.4 Transport Sector Potential

In recent years, the number of automobiles has increased dramatically in China, especially private cars, but also those for freight and passenger transport.

Compared to rail, road transport uses much more energy but is generally faster and more convenient. Energy intensity for overall transport sector increased from 19 kg oil equivalent per ton-kilometer (kgoe/t-km) to 23 kgoe/t-km between 1990 and 1997.

Railway transport in China has been undergoing a transition from steam to diesel and electric locomotives, which has greatly improved energy efficiency because diesel and electric locomotives are 30 and 25% thermally efficient, respectively, compared to 8% for steam locomotives. This shift decreases the energy needed to move passengers and freight by approximately three-fourths. Nonetheless, China's energy consumption per ton-kilometer of freight is still 60% higher than that in the United States and 30% higher than in Germany.

Improved energy efficiency in road transport will derive from using more new automobiles and improving the transportation management system. These trends are already starting to emerge in China in freight transport, with the energy use per ton-kilometer decreasing 15–20% since 1980. For passenger transport, however, energy intensity (measured as energy per passenger-kilometer) has actually increased during the past two decades as a result of lower load factors for public buses.

3.2.5 Agricultural Sector Potential

As China continues to modernize the agriculture sector, mechanization will result in increased energy use. Nevertheless, significant potential for increasing energy efficiency can be achieved with more efficient water-pumping and agricultural equipment. The Chinese government is generally emphasizing rural energy development as supplying more clean and convenient energy to farmers but not the efficient use of that energy.

3.3 Africa

The major challenge for energy development in Africa is how to increase most Africans' access to modern energy services while taking into consideration the economic costs and environmental impacts of these energy sources. Africa has vast energy resources but extremely limited energy and industrial development and the lowest rates of access to electricity in the world.

3.3.1 Trends in Energy Intensity and Consumption per Capita

As mentioned previously, energy consumption in Africa is among the world's lowest, although energy

TABLE IV
Summary of Energy Efficiency Potentials for China and Southeast Asia

Sector	Technological options/area	Southeast Asia 2020 (%)	China 2010 (%)
Industry	Fertilizer	20	10–20
	Iron and steel		15–25
	Nonferrous metals		3–16
	Cement		10–20
	Pulp and paper		20–40
	Electric motors		15–20
	Textiles		15–28
	Aluminum		5–20
	Refineries		5–10
	Brick-making		17–32
	Foundries		8–14
	Ethylene		10–30
	Calcium carbide		10–22
	Sulfate		14–25
	Caustic soda		10–30
	Glass		22
	Ceramics		30
Residential	Electric appliances	20–60	
	Illumination	20–60	
	Lighting		10–40
	Refrigerators		10–15
	Air-conditioning		15
	Washing machines		15
	Cooking appliances		20–40
Transport	Space heating appliances		10–30
	Total transport	2275 PJ	
	Two- and three-wheelers		10–15
	Cars		5–15
	Trains (diesel)		8–14
Agriculture	Trains (electric)		10
	Pump sets		20–50
	Motors		10–30

intensity is relative high. Despite the low base, primary energy consumption per capita has increased by less than 1% per year during the past two decades (Fig. 11). At the same time, energy intensity has actually increased, so African countries need more energy to produce \$1 of GDP (Fig. 12). This is true even considering the purchasing power parity difference between Africa and the rest of the world. Data for the continent, however, obviously hide vast differences between countries (Fig. 13).

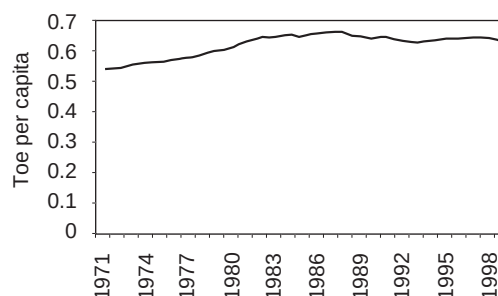


FIGURE 11 Energy consumption per capita in Africa.

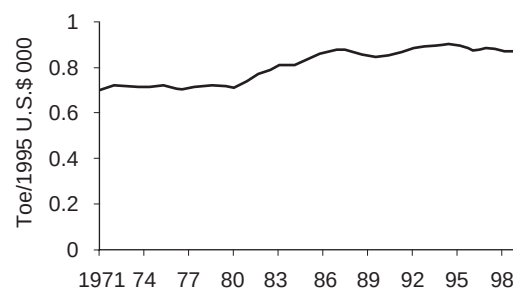


FIGURE 12 Energy intensity trends in Africa (energy use per unit GDP).

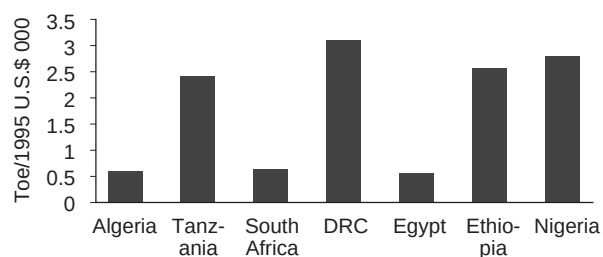


FIGURE 13 Energy intensity in African countries (energy use per unit GDP).

Although a number of African countries have included energy efficiency as part of their national energy policies, few have taken concrete steps toward legislation or specific programs and measures to promote it. Only Tunisia and Ghana have implemented mandatory appliance standards, whereas South Africa is implementing a mandatory commercial building code and has a voluntary residential building energy guideline. Ghana has created an independent body with multistakeholder support called the Energy Foundation, which has been highly successful in implementing energy efficiency programs in a range of sectors. Egypt is also considering a similar path and has set up an energy efficiency council to help develop a national energy efficiency strategy.

3.3.2 Industrial Sector Potential

Many African countries have relatively small industrial sectors, with the exception of large resource extraction industries such as mining and oil production. Even those countries, such as South Africa and some in the north, that have larger industrial sectors tend to rely on heavy, energy-intensive industrial subsectors and processes. Moreover, several southern African countries' dependence on coal, which tends to have a lower conversion efficiency in industrial boilers and power stations, also implies significant opportunities for cost-effective energy savings. An example of this is shown in Fig. 14, which compares the specific energy intensity (i.e., energy use per unit of sectoral output) of Zimbabwe to benchmark international energy intensity levels in industrialized countries.

Key industrial end uses for which industrial energy efficiency can be improved include electric motors, compressed air, lighting, process heating (boilers), and efficiency heating, ventilation, and air-conditioning (HVAC) systems. A wide range of studies suggest that the economic potential for energy efficiency in the industrial sector is 20–30%. In other words, achieving these savings is possible without increasing the cost of production, given current energy prices and technology availability.

Another useful set of examples derives from analysis of major industries in South Africa, in which better compressed air management, lighting, boilers, and changing to variable-speed drives can save 20% of the energy use for those end uses at paybacks of less than 5 years.

An important area that is not covered in the summary table for Africa is the possibility of fuel switching, particularly from coal or oil to natural gas

for industrial boilers. As Table V shows, gas boilers are considerably more energy efficient. Although electrical process heat is efficient at the point of use, the large losses in power stations mean that the overall system efficiency is closer to 25%.

3.3.3 Residential and Commercial Sector Potential (Buildings and Appliances)

The majority of households in Africa rely primarily on biomass for energy services, particularly cooking. Given that traditional use of wood and charcoal for cooking is only 15–20% efficient, whereas kerosene, LPG, and electric stoves are approximately 40, 55, and 65% efficient, respectively, an important intervention is to move households to commercial fuels for cooking. In addition, improved biomass stoves can increase efficiencies to 25–30%. The problem with assessing the economics of these options is that traditional fuel wood is generally collected and thus outside the commercial energy market. Even purchasing an improved biomass stove costs money, and purchasing fuels for stoves adds even more to household energy costs. Nevertheless, kerosene is preferred over fuel wood in most areas where it is available, and a number of successful programs to increase the use of LPG and electricity for cooking have been implemented, such as in Senegal, Botswana, and South Africa.

Where electric appliances are used in the home or for commercial applications, the potential for savings is significant. Switching to compact fluorescent lighting, for example, can save two-thirds of the lighting energy use, and proper passive solar design of houses along with ceilings and insulation (especially in low-cost housing) can save 60–70% of space-heating energy from a variety of sources. Commercial energy intensity, as with industrial energy use, tends to be high, with many opportunities for improving the efficiency of HVAC systems, lighting, and office

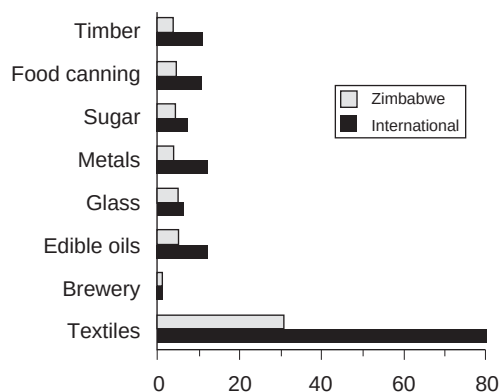


FIGURE 14 Examples of relative energy use efficiencies in industry (GJ/ton).

TABLE V
Energy Efficiency of Boilers by Fuel

Fuel technology	Overall efficiency including steam system losses (%)
Electrical process heating	75
Coal-fired boilers	60
Oil-fired boilers	65
Wood-fired boilers	60
Gas-fired boilers	70

equipment. Many housekeeping measures can save energy with almost no capital cost.

3.3.4 Transport Sector Potential

As with industry, South Africa and most north African countries have much higher levels of motorization than other countries in Africa. Road transport is the main mode in Africa. Most vehicles are imported and generally old and poorly maintained, which decreases their efficiency, as does their operation in congested urban areas and on poor-quality road networks. Improving the quality of the road network, maintaining vehicles properly, and shifting road transport to rail could all significantly improve the energy efficiency of transport. Providing affordable, safe public transport systems would also shift passenger travel away from road traffic and increase energy efficiency as well as reduce traffic congestion. The estimates in Table VI are based on a combination of shifting to public transport and improved road and rail transport efficiency.

3.4 Latin America

During the last decade of the 20th century, the energy sector in Latin America underwent profound changes. In many countries, formerly state-owned monopolies were privatized, electricity and gas sectors restructured, and energy markets liberalized. Foreign direct investment in the energy sector in Latin America has been higher than in any other region in the world. Interconnections between countries are increasing and the market penetration of natural gas is changing the energy mix in many countries. Although energy prices are declining due to competition and alternative sources of commercial energy, many of the notorious problems of energy supply in the region have remained as serious and urgent as they were before the reform process started. Among these problems are (i) disparities of energy supply in more developed urban and less developed rural areas, (ii) insufficient access to clean and convenient energy services by the urban and rural poor, (iii) poor quality of service and protection of consumers, and (iv) only marginal attention to energy efficiency and decentralized renewable energies in official mainstream energy policies. The efficient use of energy in the whole supply chain, from energy extraction and transformation to the end user, could be drastically improved to allow for the expansion of energy services while safeguarding sustainable economic and social development.

TABLE VI

Summary of Energy Efficiency Potentials for Africa

Sector	Technological options/area	Economic potential (% of current use)
Industry	Total industry	15–35
	Industry, compressed air	>20
	Industry, variable-speed drives	>20
	Industry, efficient motors	>1–5
	Industry, efficient HVAC	>10
	Industry, efficient boilers	15–20
	Iron and steel	7
	Cement	10–15
	Aluminum (secondary)	45
	Refineries	6
	Inorganic chemicals	19
	Consumer goods	25
	Food	16–30
	Cogeneration	20–25
Residential	Total energy	5–20 ^a
	Electric appliances	11–25
	Lighting, compact fluorescent lights	50–70
	Refrigerators	30
	Space heating	60–70
Commercial/public	Commercial air-conditioning (heat and cool)	50
	Commercial lighting	10
	Computers	10
		20–25
Transport	Cars, road system	30
	Total transport	30
	Road transport	30

^a Market potential.

3.4.1 Trends in Energy Intensity and Consumption per Capita

Energy intensity in Latin America is substantially higher than in industrialized countries but lower than that of many developing regions (Fig. 4). At the same time, energy consumption per capita is only one-fourth that of industrialized countries (Fig. 1). As shown in Figs. 15 and 16, the energy intensity of Latin American economies has remained quite constant during the past 30 years, whereas the total primary energy supply per capita has increased only 30% during this period, with stagnation during the 1980s and early 1990s. Both energy intensity and per capita consumption differ substantially among different countries of the region (Fig. 17).

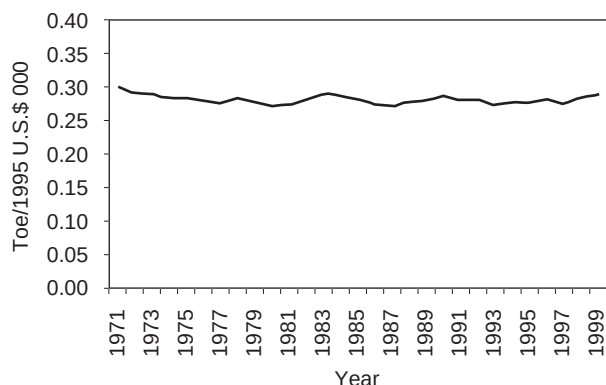


FIGURE 15 Energy intensity in Latin America.

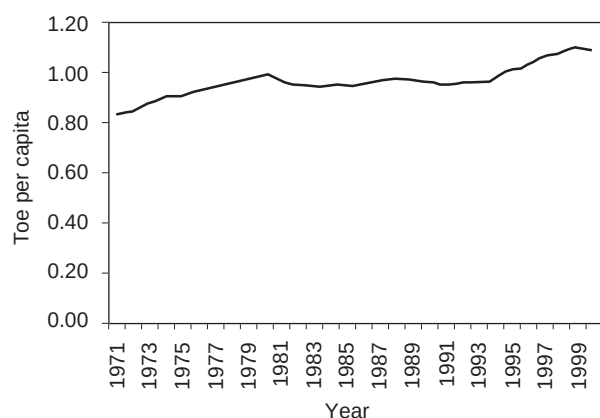


FIGURE 16 Primary energy consumption per capita in Latin America.

The use of noncommercial energy such as fuel wood is still high, particularly in rural areas, whereas low-grade and polluting fuels, such as charcoal and kerosene, are widely used in low-income urban areas. In Peru, for example, 63% of the energy consumption in the residential and commercial sectors in 1995 was fuel wood, 13% kerosene, and 6% animal dung compared to 9% electricity, 6% LPG, and 0.1% natural gas.

During the 1980s and 1990s, energy efficiency programs and legislation were developed and implemented in various countries of the region. Pioneers in this area are Brazil and Mexico, followed in the early 1990s by Costa Rica and later by other countries, such as Argentina, Colombia, Chile, Ecuador, and Peru. These energy efficiency programs were financed partly from own national sources but also with contributions from international donors. Although foreign funds have been crucial, the lack of commitment of national governments has in some cases impeded the wider uptake of the energy

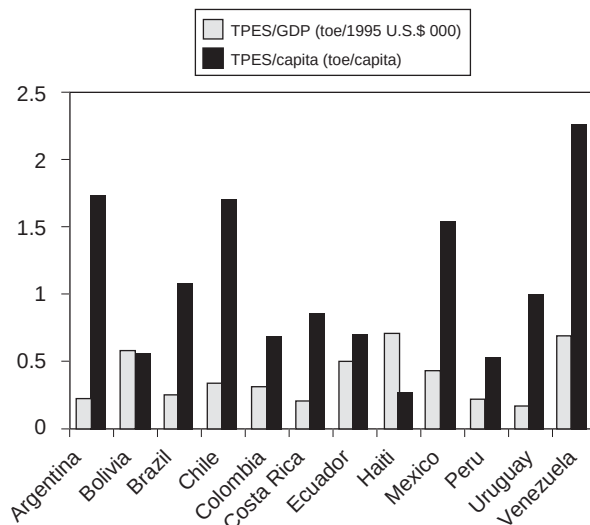


FIGURE 17 Energy intensity and per capita consumption in various Latin American and Caribbean countries, 1999.

efficiency message by the market and society. Brazil, Colombia, Costa Rica, and Peru have introduced new specific legislation for the promotion of energy efficiency and renewable energies, and bills for energy efficiency laws have been presented in other countries. These energy efficiency laws basically define the authority of governments to act in specific areas, such as information dissemination, energy efficiency standards and labeling, mandatory energy audits, the promotion of energy service companies or fiscal incentives. For example, although the elaboration of technical standards and energy efficiency labeling schemes is gaining momentum in various Latin American countries, important and visible schemes have been implemented only in Brazil and Mexico. Although the application of the standards in Mexico is mandatory and includes minimum levels of energy efficiency, the labeling scheme applied in Brazil has been voluntary. However, mandatory minimum efficiency standards are under serious consideration following the energy crisis in 2001 (Table VII).

3.4.2 Industrial Sector Potential

The market opening during the 1990s has forced many industries in Latin America to modernize their production processes in order to compete with foreign companies in domestic and international markets. Foreign direct investment has taken place in many industrial sectors and substantial technology transfer has been achieved. The modernization of industrial processes often involves higher energy efficiency and fuel switching. An example is the

TABLE VII
Summary of Energy Efficiency Potentials in Latin America

Sector	Technological options/area	Economic potential (% of current use)
Industry	Industry, efficient motors and drives	15–30
	Industry, process heat	20–50
	Industry, refrigeration	15–40
	Industry, efficient boilers and steam systems	20–40
	Iron and steel	10–30
	Cement	10–40
	Food and beverages	20–30
	Textiles	20–30
	Cogeneration	20–30
	Total residential	20–40
Residential	Electric appliances (general)	20–40
	Lighting, compact fluorescent lights	30–50
	Refrigerators	35–50
	Space heating	25–50
Commercial/public	Total commercial/public	15–40
	Air-conditioning	20–30
	Commercial lighting	30–40
	Public lighting	20–45

large-scale substitution of pyrometallurgical processes in the Chilean copper industry by modern, more efficient electrochemical processes. Whereas large industries are successfully adapting to more efficient processes, small- and medium-scale industries in Latin America often face many problems adapting to the more competitive environment.

The energy-intensive sectors cement, iron and steel, chemicals, and food and beverages consume approximately 60% of industrial energy in Latin America. Several studies carried out during the 1990s estimate that the energy conservation potential in these activities is 10–30%. The same range of savings potential also applies for the textile, other nonmetallic minerals, and machine building industries, with the majority of measures having favorable payback periods. Electrical motors and drives, process heat, and refrigeration are important areas of industrial energy end use. The energy savings potential in the field of electrical motors and drives is up to 30%, depending on the kind of measures taken. Losses in industrial thermal processes are as high as 70%, and energy conservation potentials in industrial steam

systems are as high as 40%. There is also a significant potential for industrial cogeneration in Latin America, which is largely untapped because of unfavorable legal frameworks.

3.4.3 Residential and Commercial Sector Potential (Buildings and Appliances)

Residential, commercial, and public buildings offer high energy savings potentials throughout Latin America. Building codes, where existing and enforced, do not normally include requirements concerning thermal insulation. This is particularly relevant for housing for the poor, where low-cost materials and designs are usually applied. Forthcoming standards for energy certification of residential, commercial, and administrative buildings in Chile are an encouraging step in the right direction.

Another area with major energy savings potential is household appliances. The widespread use of low-grade fuels, such as kerosene, fuel wood, and charcoal, in rural and poor urban households leads not only to high levels of pollution but also to high energy losses (up to 90%). Studies indicate that the economic potential for energy efficiency improvements in residential end uses such as cooking, lighting, and refrigeration is 20–50%. On a national basis, the Costa Rican government estimates that the combined energy savings potential of efficient refrigeration and lighting systems in the residential and commercial sectors is 12–22% until 2015, taking into consideration different levels of efficiency and market penetration of the equipment. Similar potentials for end-use efficiency (typically 20–40%) exist in commercial and public buildings (e.g., shopping centers and hotels) and in street lighting.

Social problems, such as poverty and unequal distribution of income, are an important barrier to a more rigorous transformation of the residential end-use sector. Attention to energy efficiency is dampened because of more urgent, basic service needs, and new, efficient equipment and housing are often not affordable to low-income households. Several projects in the region funded by the GEF focus on end-use efficiency in the residential, commercial, and public sectors, particularly on efficient residential and street lighting.

3.4.4 Transport Sector Potential

Road transport is the most important transport mode in Latin America, with problems such as traffic congestion, air pollution in cities, and poor fuel economy due in part to the age of the car and truck

fleet. Effective transportation systems combined with supportive urban planning, building on the model originally established in Curitiba, Brazil, in the 1970s, and the increasing use of compressed natural gas and LPG in passenger and goods transport in Argentina, Brazil, and other countries are examples of measures taken to mitigate this situation. Nevertheless, energy use in the transport sector is growing and will further grow with increasing welfare and consumption levels. Measures to reduce the energy intensity of the transport sector would require improving the quality of roads, implementing effective regulation to maintain vehicles properly, as well as incentives to use cleaner and more efficient fuels and cars, introducing effective transport systems in more cities and revitalizing railway transport in Latin America. Energy savings potential due to more efficient driving habits and preventive maintenance have been estimated to be 15–25% of the fuel consumption in passenger cars, trucks, and buses.

SEE ALSO THE FOLLOWING ARTICLES

Cultural Evolution and Energy • Development and Energy, Overview • Energy Efficiency, Taxonomic Overview • Energy Ladder in Developing Nations • Global Energy Use: Status and Trends • Indoor Air Quality in Developing Nations • Industrial Energy Efficiency • National Energy Policy: Brazil • National Energy Policy: China • National Energy Policy: India • Obstacles to Energy Efficiency • Sustainable Development: Basic Concepts and Application to Energy • Women and Energy: Issues in Developing Nations

Further Reading

- Bhagavan, M. R., and Karekezi, S. (eds.). (1992). "Energy Management in Africa." Zed Books, London.
- China International Engineering Consultant Corporation. (1997). "Comparison between China and Foreign Countries on Unit Energy Consumption of Energy Intensive Industries." State Development Planning Commission, Beijing.
- Das Gupta, M., and Roy, R. (2000). Manufacturing energy use in India: A decomposition analysis. *Asian J. Energy Environ.* 1(2), 223–250.
- Das Gupta, M., and Roy, J. (2002). Estimation and analysis of carbon dioxide emissions from energy intensive manufacturing industries in India. *Int. J. Energy Environ. Econ.*
- Geller, H., Jannuzzi, G., Schaeffer, R., and Tolmasquim, M. T. (1998). The efficient use of electricity in Brazil: Progress and opportunities. *Energy Policy* 26(11), 859–872.
- International Energy Initiative (2001). *Energy Sustainable Dev.* 4(4). [Special issue on the work of the Working Group on Energy Strategies and Technologies of the China Council for International Cooperation on Environment and Development]
- Jochem, E. (2000). Energy end-use efficiency. In: "World Energy Assessment: Energy and the Challenge of Sustainability" (United Nations Development Program, United Nations Department of Economic and Social Affairs, and World Energy Council, Eds.), pp. 173–217. United Nations Development Program, New York.
- Karekezi, S., and MacKenzie, G. A. (1993). "Energy Options for Africa: Environmentally Sustainable Alternatives." Zed Books, London.
- Liu, Z., Sinton, J. E., Yang, F., Levine, M. D., and Ting, M. (1994). "Industrial sector energy conservation programs in the People's Republic of China during the Seventh Five-Year Plan (1986–1990), Report LBL-36395." Lawrence Berkeley Laboratory, Berkeley, CA.
- Ministry of Environment and Energy (XXXX). "Programa Nacional de Conservación de Energía 2001–2006." National Commission for Energy Conservation, San Jose, Costa Rica.
- Ministry of Mines and Energy. (1998). "Plan de Eficiencia Energética para el Mediano y Largo Plazo, 1999–2009." Ministry of Mines and Energy, Lima, Peru.
- Mongia, N., Sathaye, J., and Mongia, P. (1994). Energy use and carbon implications in India: Focus on Industry. *Energy Policy* 22(11), 894–906.
- OLADE/CEPAL/GTZ (1997). Energy and Sustainable Development in Latin America and the Caribbean: Approaches to Energy Policy. OLADE/CEPAL/GTZ, Quito, Ecuador.
- OLADE/European Commission (1995, January). Proyecto "Conservación de Energía en la Industria: Esquemas Subregionales de Financiamiento."
- Roy, J., Sathaye, J., Sanstad, A., Schumacher, K., and Mongia, P. (1999). Productivity trends in Indian energy intensive manufacturing industries. *Energy J.* 20(3), 33–61.
- Sathaye, J. A., and Ravindranath, N. H. (1998). Climate change mitigation in the energy and forestry sectors of developing countries. *Annu. Rev. Energy Environ.* 23, 387–437.
- Spalding-Fecher, R., Clark, A., Davis, M., and Simmonds, G. (2003). The economics of energy efficiency for the poor: A South African Case Study. *Energy Int. J.* 27(12), 1099–1116.
- Tata Energy Research Institute. (1999). "TERI Energy Database Directory and Yearbook 1998–99." Tata Energy Research Institute, New Delhi.
- Thorne, S. (1995). "Energy Efficiency Potential in the South African Economy: A Review." Energy and Development Research Centre, University of Cape Town, Cape Town, South Africa.
- Trikam, A. (2002). Greenhouse gas mitigation options in the industrial sector. *South African J. Econ. Management Sci.* 5(2), 473–498.
- Winkler, H., Spalding-Fecher, R., Tyani, L., and Matibe, K. (2002). Cost-benefit analysis of energy efficiency in urban low-cost housing. *Dev. Southern Africa* 19(5).

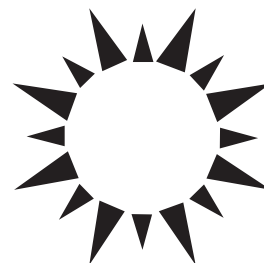
Prices of Energy, History of

PETER BERCK

University of California, Berkeley
Berkeley, California, United States

MICHAEL J. ROBERTS

U.S. Department of Agriculture¹
Washington, D.C., United States



-
1. The Long Historical Sweep
 2. The Cost of Finding Oil
 3. The Organization of Petroleum Exporting Countries and the Oil Embargoes
 4. Coal Prices
 5. Natural Gas Prices
 6. Energy Prices to Users
 7. A Market for “Energy”: Cointegration of Prices
 8. Prices and the Nonrenewable Resource Price Theory
 9. Conclusion: Supply Factors and Demand Factors

Glossary

capital asset pricing model Predicts that an asset's risk-adjusted rate equals the risk-free rate plus an asset's beta multiplied by the expected return on the market portfolio (aggregate wealth). An asset's beta is its covariance with the market portfolio divided by its variance.

cointegration The property that a linear combination of data series is stationary while the individual series is not stationary.

marginal cost The cost of producing one additional unit of a good.

stationarity The property that the expected mean and variance of a series do not change over time.

The earliest energy product for which there are prices is wood. Wood has other, and higher valued, uses than as a source of energy. It is also expensive to transport per British thermal unit provided, hard to burn efficiently, nonuniform in composition, and

¹The views expressed are those of the authors and do not necessarily reflect the views or policies of the U.S. Department of Agriculture.

very high in potentially polluting by-products. For these reasons, it has been supplanted in the energy-intensive developed world. However, it remains an important source of fuel in less developed countries. Coal rapidly supplanted wood as the fuel of the industrial revolution, and as a product of British society (and its American offshoot) there are long series of data available on the price of coal, or coales as it was first called. The percentage share of coal in the energy market has decreased for reasons of transport and environment. Efficient transport for coal is now provided by dedicated trains (unit trains), although water transport, including canals, was the principal method until the late 19th century. The need for expensive fixed transport has made coal hostage to railroad pricing policies and to government regulations. Water power was also an important industrial energy source in the early industrial era, but it was harnessed in a local fashion (mills on streams), and so we are unaware of price series for this direct use. Later, of course, water was used to produce electricity, and there are ample records of electric prices and estimates of hydropower costs. Nuclear power is likewise used to manufacture electricity, although it has been used only since the late 20th century. Price records are for electricity, regardless of its source, although cost estimates are available for nuclear power. The gas industry began with manufactured gas, usually derived from coal, which was used for lighting. In the United States, Baltimore was the first city to have piped gas in 1816. The transmission of natural gas began in 1872, but the 8-in. pipeline from Greentown, Indiana, to Chicago in 1891 was the first large commercial natural gas venture. Natural gas, lacking the carbon monoxide of manufactured gas, was suitable for heating. Natural gas is now an important source for

heating and power generation. Its emission profile is more favorable than that of coal or oil, which helps account for its growing usage. Price series for natural gas are not nearly as long as for coal because large-scale commercial use only began in the 20th century. Oil is the last fuel mineral that we consider. The first oil well was drilled in 1859 near Titusville, Pennsylvania, and oil rapidly became the fuel mineral of choice. Oil burns cleaner and is easier to transport than coal. It is also easier to use because oil furnaces have no need for ash removal or other hand work. Oil price history is well recorded. Because of Standard Oil, the Texas Railroad Commission, and the Organization of Petroleum Exporting Countries, oil prices are part of modern political debate and consciousness.

1. THE LONG HISTORICAL SWEEP

Figure 1 presents Hausman's real price index series for wood and coal for Great Britain from 1450 to 1989. Hausman normalized the index to 100 in 1600 in the belief that by then coal was cheaper than charcoal so that the figure would understate the advantage of coal over wood as a fuel. There is not sufficient evidence concerning heat content to normalize the two series in that fashion. The series are meant to capture the price of the fuel source and are net of transportation or taxes. The figure clearly shows the increasing price of wood and decreasing price of coal in the 1600s. In fact, coal does not begin to rapidly increase in real price until nearly the

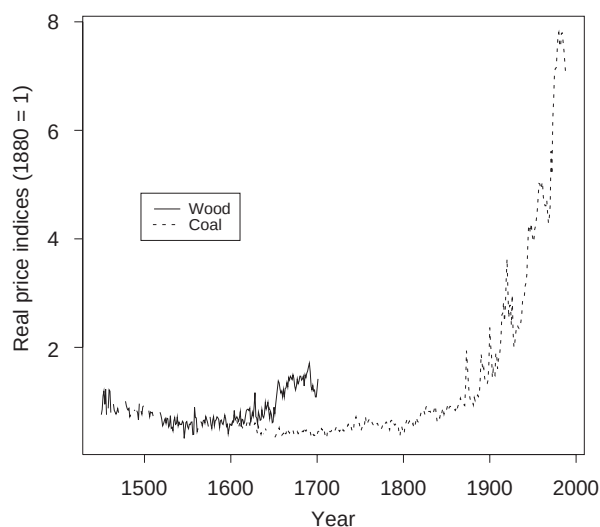


FIGURE 1 Historical prices for Great Britain fuel—wood and coal. Adapted from Hausman (1995).

beginning of the 20th century. In addition to cartels, strikes, and recessions, which account for some of the peaks and dips in Fig. 1, the major change in the U.K. coal industry was nationalization after World War II. The overall increase in 20th-century price is due to the quality of U.K. coal deposits and the depletion of more favorable seams. This is peculiar to the United Kingdom.

Figure 2 shows the prices of the dominant energy minerals in the United States from 1870 to the present. The prices are in real 1970 dollars per British thermal unit (Btu). The price of oil decreased significantly from 1870 to the energy shocks of the 1970s and 1980s and then returned to prices not unusual in the early 20th century. A more complete discussion of the energy shocks is provided later. Coal prices are less than those of oil, reflecting oil's greater portability and cleaner burning. Coal prices increased along with oil prices after the 1970 energy price increases. Natural gas was regulated until 1979 and then deregulated during the next decade. Particularly after 1970, the clean burning characteristic of natural gas made it a desirable fuel. This property and the end of regulation account for the change in the relative prices of coal and oil.

Figure 3 shows U.S. energy consumption by source, and it indicates the increasing importance of natural gas, oil, and coal and the decreasing

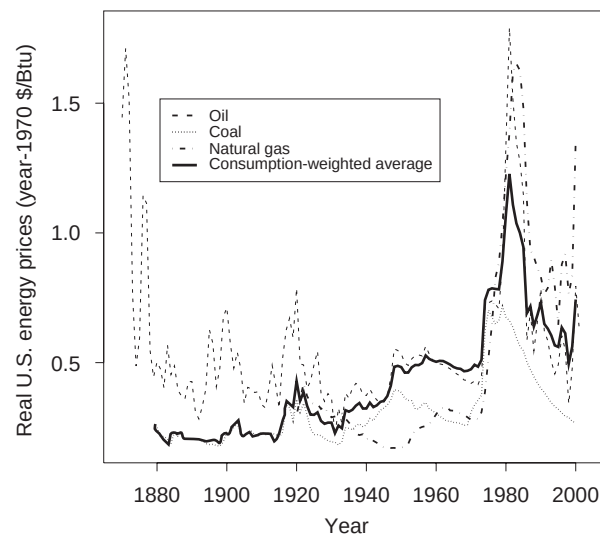


FIGURE 2 Historical prices in the United States, 1870–2001. Annual data on current prices for the raw resource products were obtained from the Energy Information Administration (EIA) and converted to 1970 dollars using the consumer price index. These values were converted to approximate dollars per Btu based on the following data from the EIA: In 1970, electric utilities paid on average approximately 28¢ per million Btu of natural gas, 31¢ per million Btu of coal, and 42¢ per million Btu of petroleum.

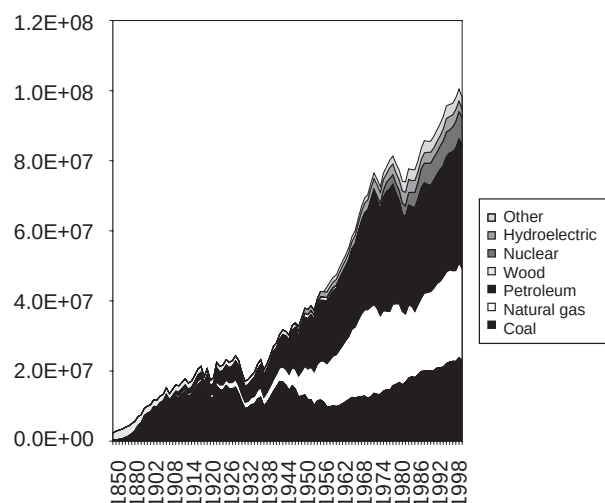


FIGURE 3 U.S. energy consumption by source (trillion Btus). Data from the Energy Information Administration and historical statistics.

importance of wood. Coal was the first energy mineral to produce significant energy output. Coal energy usage increased slowly during the 20th century. At the beginning of the century, coal accounted for approximately 73% of U.S. energy use, with wood accounting for nearly all the rest. Today, coal, gas, and petroleum account for approximately 22.6, 24, and 39.4% of use, respectively, with the remainder coming from renewable sources, primarily nuclear and hydropower.

2. THE COST OF FINDING OIL

The private cost to the economy of using a barrel of oil is the cost of finding a barrel of new reserves of the same quality to replace the used-up barrel. If the oil business were competitive, which it is not, then the market price of oil would be the finding costs plus the development costs plus the marginal costs of production. For this monopolistic business, it is the cost of finding oil that provides the most information about oil's ultimate scarcity. The cost series for finding oil are necessarily very incomplete and lack certainty. The series developed by the Energy Information Administration (EIA) comprise expenditures on exploration and development divided by oil and gas reserve additions. The latter series is subject to considerable uncertainty since for many years it is unknown how much oil (and gas, valued as oil equivalent) was actually found. Additions to reserves can be dominated by changes in the estimates of existing fields, whereas it is the cost of developing

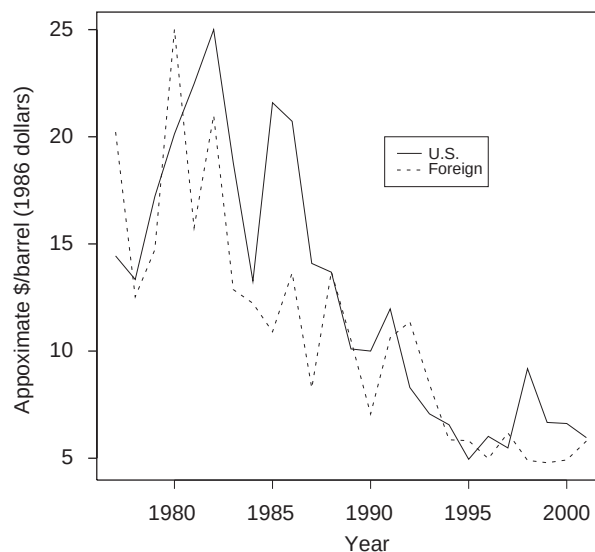


FIGURE 4 Cost to find oil and gas. Estimates for the cost to find oil were derived from data obtained from the Energy Information Administration, Form EIA-28 (Financial Reporting System). The left axis represents total expenditures on exploration and development divided by new additions to reserves. Total reserves include both oil and natural gas. Natural gas reserves were converted to barrels using the average September 1998 prices for natural gas and oil.

new sources that is of most interest. Information was collected for the years 1977–2001 for both foreign and domestic activities of U.S. reporting firms and is plotted in Fig. 4. For 1977, the cost was calculated in 1986 dollars as \$14/barrel U.S. and \$20/barrel foreign. Domestic costs peaked at \$25 in 1982, declined to \$4.95 in 1995, increased again, and then settled at \$6/barrel in 2001; foreign costs mostly declined during the period and ended at \$6. It is believed that the cost of finding and developing oil in the Persian Gulf is less than these figures.

3. THE ORGANIZATION OF PETROLEUM EXPORTING COUNTRIES AND THE OIL EMBARGOES

On October 17, 1973, the Organization of Petroleum Exporting Countries (OPEC), which was originally formed in 1960, announced a progressive embargo on the sale of oil to the United States and other countries that supported Israel in the 1973 Yom Kippur war. They also announced a 70% increase in the reference price for crude oil, an artificial price that determines how much oil companies must pay in taxes. Given that the United States bought only 5% of its oil from

OPEC, an embargo of the type announced was unlikely to inconvenience the United States or substantially raise prices. However, Saudi Arabia decreased its production 35% below the September output and increased the price an additional 6¢. Combined with the nationalization of oil company interests, the cutbacks in production resulted in major increases in the price of oil and a major redistribution of the rents from oil production to the oil-producing companies. In retrospect, it is clear that these actions by Saudi Arabia, with some support from the rest of OPEC, were largely commercial in nature and U.S. attempts at bilateral relationships and other diplomacy had no effect. Production limitation became an effective tool for raising prices far above discovery and production costs.

The next increase in prices occurred in 1978–1981 as a result of a secret OPEC program to impose production controls and due to turmoil in Iran, including the Iran–Iraq war. When Iran's oil output decreased, Saudi Arabia did not increase its output. With Iran not producing at all in January 1979, Saudi Arabia decreased production by 2 million barrels per day on January 20. Its output vacillated for the next several months, always well below capacity, with the result that oil prices topped \$20/barrel in October 1979. Prices increased until 1982, when the Saudi export price reached \$33.50/barrel. In this period of war-induced uncertainty, oil companies and oil users hoarded significant amounts of oil.

Adelman attributes the price decline in 1982 to the end of a speculative bubble. The continued slow decline until 1986 was the result of increased supply from non-OPEC members and the consequence of lagged demand elasticity. In 1986, prices declined considerably, presumably a function of the substantial non-OPEC output. The lowest price was reported by Adelman as \$6.08 for Saudi Light; in August of 1986, OPEC again set production quotas, with cooperation from some non-OPEC members.

The next major price event was the Gulf War in 1991, again producing a spike in prices. The war directly ended the sales by Iraq and Kuwait and spurred the building of precautionary supplies. Because there was sufficient capacity without the output of these two countries, prices rapidly subsided.

4. COAL PRICES

In the time of Queen Elizabeth I, the coal trade in England (i.e., coal sold into the London market) was monopolized. Monopolization was possible because

the London market was served by two rivers, the Tyne and the Wear, and the coal lands adjacent to these rivers. The land involved was not so extensive as to defy ownership by an oligopoly. The key issue in the early coal trade was water transport, with coal lands situated near canals or rivers with major markets being valuable and other lands being of lesser value.

Although oligopoly was in existence as early as 1517, it was formalized with royal charter as a monopoly of coal and also of political power in Newcastle by Queen Elizabeth on March 22, 1600. Other important dates cited by Sweeney include an act of Parliament to make restriction of trade in coal illegal in 1710, an allocation document fixing the quantities of coal among major owners in 1733, a lack of regulation in the period 1750–1771, a resumption of regulation in 1771 or 1772, and on-and-off regulation through 1850. Although the detailed history is not provided here, the British coal series should be interpreted in the light of the on-and-off nature of the quantity regulation. Price changes may reflect temporary changes in the Lerner index as much as changes in capacity or demand.

Coal prices in the United States during the 20th century were subject to a very different sort of government regulation. Because of the need to reduce sulfur emissions in the air, the U.S. government required new and substantially modified power plants to reduce sulfur emissions in such a way that there was no advantage to the use of low-sulfur rather than high-sulfur coal. In effect, since 1990 the emissions permit trading program no longer disadvantages low-sulfur coal as a way to reduce emissions.

5. NATURAL GAS PRICES

In the United States, natural gas was regulated and sold at a low price until gradual deregulation from 1979 to 1989. Thus, the prices after 1989 were free to increase in response to increased demand, driven by the desire to switch from expensive oil to gas and by environmental restrictions that allow the burning of gas but restrict oil and coal. For the data plotted in Fig. 2 prior to 1979, the correlation between gas price changes and oil price changes is 0.067; after deregulation, the correlation is 0.605.

6. ENERGY PRICES TO USERS

Although there is a world market in many of the energy minerals, national governments' tax and

allocation policies create very different prices for end users in different countries. In general, the United States has among the lowest prices of consuming nations to end users of energy products, and all countries generally price consumer products higher than industrial products.

In comparing prices across countries, one must keep in mind the problems of converting currencies and the qualities of the fuel. The following discussion is in terms of conversion at the official exchange rate as opposed to purchasing power parity. The exchange rate is the best choice for industrial uses, especially where the end products (e.g., cars, steel, and airplanes) are widely traded. Particularly in countries less well attached to the international trading system, exchange rates can be very different than purchasing power parity, and the use of dollars as the official exchange rate would provide a poor estimate of the burden on consumers of energy purchases relative to a basket of other goods. In order to make industrial and household prices comparable, we use exchange rate conversion throughout. There is also an issue of heat equivalent [or heat equivalent net of the energy needed to remove water from the fuel, gross caloric value (GCV) and net caloric value (NCV), respectively.] When comparing prices across fuels (oil versus coal) or across different types of the same fuel (U.S. bituminous coal versus Czech brown coal), the most accurate comparisons are in terms of ton of oil equivalent or NCV. In order to use tax information, we use the reported metric tons (tonnes).

Perhaps the most traded energy commodity is fuel oil. The following section reviews the evidence that fuel oil is one international market; that is, the prices in different countries increase and decrease simultaneously, with differences only in transportation costs. For 2001, the range of prices for high-sulfur fuel oil for industry, in U.S. dollars per tonne, was 103 to 258. The lowest price in the International Energy Administration's (IEA) data set was for Mexico, which is a net exporting nation. (The data do not include the Gulf States.) The U.S. price was \$147/tonne, that of France was \$154.4/tonne, and that of Turkey was \$181.7/tonne, whereas the Organization for Economic Cooperation and Development (OECD) average was \$189/tonne. France levied taxes of \$16.6/tonne, whereas the United States levied no taxes; therefore, the pretax price in France was cheaper than in the United States. Turkey's tax was \$42 and its pretax price was cheaper than that of the United States, which should not be surprising given that Turkey is a transshipment point of Mideast oil.

Automobile fuel is refined from oil. Apart from issues of taxation, the prices of automobile fuel should be very similar throughout the world, just as the ex-tax prices of fuel oil are very similar. Consumers pay the price with tax, which varies greatly from country to country. Of the countries that report a price for regular unleaded gasoline, the United States has the lowest at \$0.39/liter. Germany has a price of \$0.90 and Denmark is highest at \$0.97. France reports a premium leaded price of \$1.00 and Turkey \$0.99, and Mexico reports a premium unleaded price of \$0.65 per liter. This variance in pricing is attributable to taxation. Although it is not reported in the IEA data, there can also be a vast difference in the attributes of gasoline. In order to meet its obligations under the Clean Air Act, the state of California requires the use of a much more refined gasoline than is used in the rest of the United States and this gasoline carries a hefty premium in price.

Coal is a basic energy source both for industry and for electric generation. Coal is more difficult to transport than oil (railroads and slurries versus pipelines) and major deposits are often not near ocean transport, unlike oil. Coal is also more likely to be used without much preprocessing (for pollution control purposes, it can be washed, whereas crude oil is always refined), so coal must be matched much more closely to boilers than crude oil needs to be matched to its ultimate customers. For these reasons, there should be a greater variance in the underlying price of coal than there is in the underlying price of oil. Again examining IEA data, the prices for steam coal (for industry or electric generation as opposed to coking or metallurgical coal), the U.S. price was \$36/tonne and that of Japan was \$32/tonne, with many other countries in the \$30 range. Germany was highest with a price of \$52/tonne. Coal can also be measured in tons of oil equivalent to compensate for different caloric outputs per ton, and the pattern is much the same.

Electricity is generated from a variety of processes, including burning of coal, oil, and natural gas, nuclear reactors, hydropower, wind, photovoltaic, and even tide power. A country's electricity prices to industry and to consumers represent a mix of the costs of these methods. Countries such as Norway, or areas of countries such as the U.S. Pacific Northwest, that have significant surpluses of hydropower can have very low costs of production, as can countries that are willing to burn coal, particularly with minimal pollution control. Variation in electric rates across countries is thus explained by natural abundance of cheap methods, preferences for clean air, and taxation. In

2000, the U.S. price for households was \$0.082/kWh (excluding tax) and it was \$0.042/kWh for industry. Denmark had the highest household price at \$0.197/kWh, and the Slovak Republic had the cheapest at \$0.05/kWh. The Slovak price is probably related to the burning of brown coal, with its concomitant problems of air pollution. Norway's household price was \$0.057, reflective of both hydropower and petroleum reserves. Industrial prices in Europe are much lower than household prices, largely due to taxes. France's industrial price was \$0.036/kWh compared to its household price of \$0.102/kWh. Italy, Japan, and Turkey are outliers in the IEA data set because they had industrial electric prices of \$0.089, \$0.143 (1999), and \$0.080/kWh, respectively.

Natural gas for households on a toe-NCV basis also show great variation among countries, with the U.S. price at \$417, slightly above the OECD average of \$391 in 2001. This is quite different from 1994, when the U.S. price was \$274 compared to an OECD average of \$383. Presumably, this change in relative price is related to shortages of both pipeline space and gas at the wellhead. Despite the clear preference in the United States to generate electricity with natural gas for air quality reasons, we expect that the price will decrease. With a price of \$788, Denmark has the highest price in the dataset.

Taxation represents a large percentage of the delivered price of energy-containing products, particularly consumer products. Automotive fuel taxes represent half or more of the value of the product in all but 4 of the 29 countries included in the IEA dataset. In the United Kingdom, they are 76.9% of the value.

The earlier years in the IEA data set tell much the same story as the later years. Pricing to consumers tends to have more taxation and more idiosyncratic variation than pricing to industries.

7. A MARKET FOR "ENERGY": COINTEGRATION OF PRICES

Energy is not a homogeneous resource and the prices of the various energy minerals do not necessarily track each other well over time. Oil is the most transported of the energy minerals, much of it originating in the Persian Gulf and transported by tanker. The differences in the prices of oil at different ports can be explained largely by transport costs. One way to test the hypothesis that oil is traded in a single market is to determine whether the series are

cointegrated. Two price series are cointegrated if there exists a linear combination of the two series that is stationary (the mean and variance do not change over time). In other words, the two series move together.

An example of this type of test is to regress the price of Saudi Light landed in southern Europe on the price of the same oil landed in Rotterdam and then test the residuals of the regression to determine if they are stationary. Given the oil series are both stationary in differences, a finding that the residuals of this regression are also stationary means that the southern Europe price series is a linear function of the Rotterdam series, and any remaining differences between the two series can be accounted for by transport costs. It means that oil price shocks affect both series at the same time. Cointegration tests for oil at different ports find cointegration and support the notion that oil, in Morris Adelman's words, is a single large pool.

Coal contributes to world energy somewhat less than oil and so if there were one price for "energy," then oil and coal prices would also be expected to be cointegrated. Similarly, coal prices in different localities would be cointegrated. There is little evidence that coal prices move together or that they move with oil prices. Figures 1 and 2 show the prices of coal for the United Kingdom and the United States. In these long series, the price in the United Kingdom increases approximately eightfold, with the strong trend being upward. The price in the United States shows a peak in the early 1980s and then a sharp decline to its previous value. It is dubious whether the U.S. price is differenced stationary rather than just stationary. The U.K. price is clearly stationary only in differences. Thus, these two prices are not cointegrated. Similarly, there is little evidence that the German and U.K. prices are cointegrated.

Although the US coal and oil prices are not found to be cointegrated using the normal testing procedure, Fig. 2 indicates that the coal price does bear some relationship to the oil price, very notably in the price run-up in the late 1970s. One explanation for the difference in the U.S. coal and oil prices is that coal is sold on long-term domestic contracts and is not subject to supply disruptions. As a result, there is no incentive to hoard coal, whereas hoarding oil before the Gulf War was a major cause of the price increase.

There is a world price of oil, but the oil price is not closely related to the prices of coal in various countries, nor are the prices of coal closely related to each other. Energy is not one great pool and there

is no “price of energy,” just a multiplicity of prices of energy minerals.

8. PRICES AND THE NONRENEWABLE RESOURCE PRICE THEORY

There is no convincing evidence of a trend in oil, gas, or coal prices. One theory on these prices that has empirical support is that the first differences are stationary with mean drifts of zero. In this case, there would be no deterministic trend in the prices, simply random variation. The best estimate of future prices would be today's price. Another possibility is that the prices revert to a quadratic trend that is stochastic. This view would lead to a prediction of a slight increase in prices. Based on the historical evidence in the time series for energy prices, there is no good evidence for forecasts of increasing price.

Examining the oil price series in Fig. 2, it appears that oil prices were distinctly declining in the early years of exploitation, then leveled off, and since 1980 have increased. Since 1980, prices appear to have followed a quadratic trend. Since 2002, prices seem to have reverted to pre-Iraq war levels. We conclude that analysis of the price path of oil does not lead to a prediction of increased prices.

The lack of a clear trend—deterministic or stochastic—in energy prices at first seems at odds with the theory of natural resources. Theory predicts that prices for these commodities will ultimately rise at the rate of interest, a classic finding by Hotelling. Hotelling reasoned that if one extracted a barrel of oil today and put the money in the bank for 1 year, then one should have exactly the same financial outcome as if one extracted the barrel in 1 year. If this equality were not true, then one would be better off extracting all of one's oil in just one of the time periods, whereas we observe extraction in all time periods. A barrel extracted today would earn its price, $R(0)$, and after 1 year at interest rate i it would be worth $R(0)(1+i)$. This must be equal to the value extracted in 1 year's time, $R(1)$ and $R(1)=R(0)(1+i)$. This theory predicts that mineral prices will rise at the rate of interest. However, the price that is referred to is the seldom-observed price of minerals still in the ground, which is called the mineral rent. The price of minerals that we observe, P , is the mineral rent R plus the marginal extraction costs c . The theory is not so clear about minerals that have already been mined or lifted, which is the series that we have presented.

Since energy mineral prices do not increase at 7–11%, which is the normal rate of return for financial assets, economists have posited that either technical progress or a decline in grade account for the slow or nonexistent rates of appreciation of energy mineral prices.

One possibility is that there is technical progress in the mining technology; that is, that c decreases over time. Then the marginal cost of mining would decline and the value of the underlying mineral would increase. The observed price is the sum of the two prices and so could increase or decrease. However, since the extraction costs are decreasing and the underlying price is increasing, eventually the extraction costs would be insignificant and the price increases in the rent would be dominant. This process requires the rent to become dominant; however, if it is true that undiscovered, but easily discoverable, reserves are very large relative to demand, then rents will be very small and prices will not exhibit an upward trend for a very long time.

Flat prices can also be reconciled with the Hotelling theory by considering the fact that depleting a resource affects extraction. The first units extracted are presumably the least costly, with each unit becoming progressively more costly as cumulative extraction progresses. In this case, part of the opportunity cost of extracting a unit today is increased cost in the future. Rents increase more slowly than the interest rate, but the difference is exactly offset by an increase in future costs. Thus, in the long term, prices still increase at the rate of interest, even if rents do not. The evidence for grade depletion as a worldwide phenomena is thin—there are substantial high-grade coal seams in the United States, and extraction costs (and finding costs) for oil in the Gulf remain very low.

Another way to reconcile the flat price path for energy minerals with the theoretical prediction of price increase is to calculate the required interest rate using the same logic used to determine the needed rate of return for common stocks.

In most of the theoretical literature, the interest rate is taken as a fixed market price of time preference: the marginal value of a dollar received tomorrow relative to its value today, with all else remaining the same. In a world of uncertainty, however, the marginal value of returns depends not only on the timing of returns but also on uncertain contingencies. Losing money on a stock market investment is likely to be more painful if one also loses one's job but less so if growth in one's home equity more than compensates for stock market losses. The fair rate of return of one investment

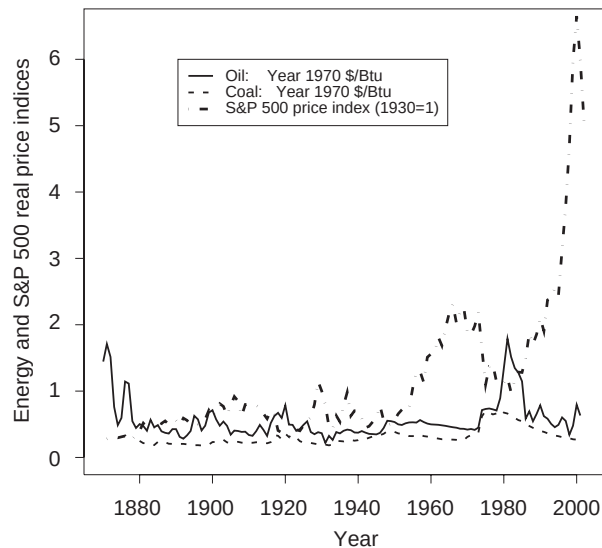


FIGURE 5 Long-term stock market returns and energy prices. The coal price and oil price data are those plotted in Fig. 2. The long historical series of the Standard & Poor's 500 was obtained from Robert Shiller (<http://aida.econ.yale.edu/~shiller/data.htm>).

therefore depends on how its returns are correlated with those of other investments.

Interest rates reported for treasury bills and other assets are typically low-risk investments that yield the same amount regardless of other contingencies. In contrast, a decision not to extract and sell a unit of oil today is an investment with a return that depends on future oil prices, which are volatile and uncertain. Thus, interest rates on treasury bills or other low-risk investments may serve as a poor proxy for the rate of return that should be applied to nonrenewable resources.

The capital asset pricing model (CAPM) explains the difference in average returns across investments according to their differential risks. The model assumes that individuals, with all else being the same, prefer greater expected returns and lower risk. To mitigate risk, individuals invest in large portfolios of assets in order to average away the idiosyncratic risk of each asset against the idiosyncratic risk of all other assets. In equilibrium, individuals thus do not care about idiosyncratic risk; the only risk that matters is that pertaining to aggregate wealth—that associated with the entire portfolio of assets.

Theory predicts that assets having returns with a strong positive correlation with wealth should have high average returns. The strong positive correlation of these assets implies that they pay off most when wealth is high and marginal utility of wealth is low. Conversely, assets with returns having a negative covariance

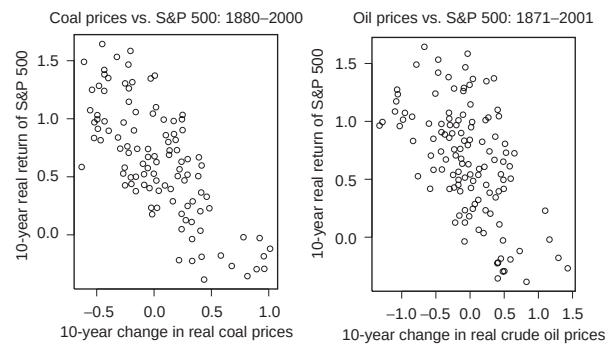


FIGURE 6 Long-term stock market returns and energy price changes. Returns on the Standard & Poor's 500 assume annually reinvested dividends. The long historical series of the Standard & Poor's 500 was obtained from Robert Shiller (<http://aida.econ.yale.edu/~shiller/data.htm>).

with wealth will earn low or even negative average returns. These assets provide insurance against aggregate losses, yielding higher outcomes when wealth and consumption are low and marginal utility is high.

Energy prices have a strong negative correlation with the aggregate economy and aggregate stock market returns and therefore should have a risk-adjusted average rate of return that is lower than that of low-risk investments such as treasury bills. The relationship between energy prices and the stock market, illustrated in Figs. 5 and 6, has received less attention but is more closely related to CAPM. The correlation between the 10-year return of the Standard & Poor's 500 and the 10-year change in coal prices is less than -0.7 . Simple CAPM-based estimates of the risk-adjusted rates of return for oil, coal, and natural gas, regardless of the time horizon, are between -1 and 2% and are not statistically different from 0% . As long as price fluctuations for these fossil fuels are strongly negatively correlated with national wealth, theory does not predict that the prices will trend up.

9. CONCLUSION: SUPPLY FACTORS AND DEMAND FACTORS

The past century of energy prices is a history of shifts in supply factors, shifts in demand factors, and policies, such as taxes and trade restrictions. However, the most notable price spikes, including those that occurred following coal strikes in the early part of the 20th century and the oil embargo, OPEC activities, and Iraq wars, are all attributable to shocks that reduced supply or anticipated supply. The observation that supply factors were most

relevant to historical price fluctuations mirrors the negative relationship that has been documented between oil prices and the macroeconomy. Because energy resources are essential inputs, energy price shocks caused by unanticipated shifts in supply cause reductions in aggregate output. If aggregate demand shocks were driving price increases then the relationship between energy prices and aggregate output and stock market returns would be positive, not negative.

The history of energy prices therefore stands in contrast to the history of energy consumption, which has grown steadily with aggregate output and demand. Over the long term, prices have trended flat and have been highly volatile, with the drivers of price shocks stemming from shifts in supply. Supply volatility has caused prices to be negatively correlated with aggregate output and wealth, which in turn explains why prices have not trended up.

SEE ALSO THE FOLLOWING ARTICLES

Business Cycles and Energy Prices • Depletion and Valuation of Energy Resources • Economics of

Energy Demand • Economics of Energy Supply • Energy Futures and Options • Inflation and Energy Prices • Innovation and Energy Prices • Oil Price Volatility • OPEC Market Behavior, 1973–2003 • Stock Markets and Energy Prices

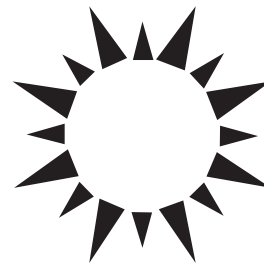
Further Reading

- Adelman, M. (1993). "The Economics of Petroleum Supply." MIT Press, Cambridge, MA.
- Adelman, M. (1995). "The Genie out of the Bottle." MIT Press, Cambridge, MA.
- Berck, P., and Roberts, M. (1996). Natural resource prices: Will they ever turn up? *J. Environ. Econ. Management* 31, 65–78.
- Hausman, W. J. (1995). Long term trends in energy prices. In "The State of Humanity" (J. L. Simon, Ed.), pp. 280–286. Blackwell/Cato Institute, Oxford, UK.
- International Energy Agency (1984–2003). "Energy Prices and Taxes." Organization for Economic Cooperation and Development, Paris.
- Livernois, J., and Martin, P. (2001). Price, scarcity rent, and a modified r-percent rule for nonrenewable resources. *Can. J. Econ.* 34, 827–845.
- Sweezy, P. (1938). "Monopoly and Competition in the English Coal Trade, 1550–1850." Harvard Univ. Press, Cambridge, MA.

Public Reaction to Electricity Transmission Lines

LUTHER P. GERLACH

University of Minnesota
Minneapolis, Minnesota, United States



-
1. Opposition to Electricity Transmission Lines Is Significant, Organized, and Informed
 2. What Is the Significance of Opposition to Electricity Transmission Lines?
 3. How Do Transmission Line Opponents Organize?
 4. What Reasons Do People Give for Their Opposition to Transmission Lines?
 5. Need: The First Defense against Transmission Line Siting
 6. "Size" of the Transmission System as a Reason
 7. Structure and Restructuring of the Electricity Utility Industry as Reasons
 8. Sentiments about Visual Impacts and Loss of Control as Reasons
 9. Health, Safety, and Electric and Magnetic Fields as Reasons
 10. Summary

Glossary

- alternating current (AC)** Flow of electricity that constantly changes direction between positive and negative sides. Most transmission lines are AC.
- direct current (DC)** Electric current that flows constantly in one direction. A significant case of opposition to a transmission line in the 1970s was to a DC system from North Dakota to Minnesota.
- electric and magnetic fields (EMF)** Invisible lines of force that surround any electrical device, such as a transmission line, electrical wiring, or an appliance. There is controversy about the health effects of exposure to these fields, especially magnetic fields.
- electricity utility** A firm established to provide electricity services. The traditional structure is one of vertical integration of generating, transmitting, and distributing services. The new structure in many countries separates these functions.

opposition group People organized to contest facility siting and building. They are usually organized in local groups interacting with others in communication networks. Groups usually include persistent activists as cadre and episodic participants in support.

siting authority The government entities authorized to make decisions about siting.

siting process The process of permitting or approving transmission or other facilities, including determining public need for them and their location or route.

transmission lines Heavy wires that carry large amounts of electricity over long distances from a generating station to places where electricity is used. The lines are held high above the ground on tall structures called towers in the United States and pylons in the United Kingdom.

Opposition is the most significant public response to transmission lines in Western industrial democracies. The study of this opposition reveals how people contest and resist transmission line siting by organizing in groups and networks, and by developing and communicating arguments about why the line is not needed and how it negatively impacts people and environment.

1. OPPOSITION TO ELECTRICITY TRANSMISSION LINES IS SIGNIFICANT, ORGANIZED, AND INFORMED

In the United States, the United Kingdom, and Canada and probably in other Western, industrially advanced democracies, the most observable and significant response of people directly affected by the siting and building of electricity transmission lines in their communities is their opposition to these projects. This article examines this public response of

opposition. It considers how people organize to oppose and examines the main reasons that people and groups give for their opposition. Major reasons given for opposition are as follows: (1) need and alternatives, (2) size/length and voltage, (3) structure of industry, (4) sentiments, visual impacts and loss of control, (5) health and safety risks, particularly from electric and magnetic fields, (6) fairness, (7) environmental justice, (8) environmental quality, and (9) siting process. This article examines the first five reasons and mentions the others. These reasons constitute a system of interdependent ideas. As people oppose the siting of transmission facilities, they produce, learn, and communicate these reasons. People build local groups to contest a particular transmission project and build global networks to spread and exchange information. The article describes this organization, illustrating communication with examples of the flow of information about electric and magnetic fields.

2. WHAT IS THE SIGNIFICANCE OF OPPOSITION TO ELECTRICITY TRANSMISSION LINES?

People can respond in several ways to the information that an electricity utility proposes to site (locate and build) transmission facilities (line and/or substations) in or near their communities. They can actively oppose the line, passively accept it, or actively support it. This article is mostly about active opposition, also termed resistance or protest. It is about opposition in the 1970s in United States and from the 1990s through the time of this writing, 2003, in the United States, Canada, and the United Kingdom. Examples from case studies in these countries illustrate opposition. This section on significance introduces many of the examples used throughout this article and such examples are indicated by an asterisk.

Findings for this article were obtained by the author, using field research, conducting a document and media study, and, for cases since the 1990s, corresponding by e-mail and gathering information from the Internet. Sources include representatives of the main parties to transmission controversies: electricity utilities and trade associations and their consultants, government siting authorities, and especially groups opposing transmission siting. Efforts to obtain information from electricity utilities, trade associations, and citizen groups on the European continent, Japan, and India through e-mail inquiries

were unsuccessful. Thus, their stories are not represented in this article. This article presents the ideas of people, especially of line opponents, but does not evaluate their accuracy. The article is about what people say and do, not what they should say and do.

This article is about opposition because this is the most observable significant response. Utility executives, consultants, government officials, and energy researchers state that opposition to the siting and constructing of transmission facilities has been one of three main obstacles to the construction of enough such facilities to meet their need and avert what some refer to as congestion of transmission. Many of these sources state that a second obstacle is the way local and state rules and processes to regulate and approve transmission facilities delay or make approval uncertain. Many also state that a third obstacle is the uncertainty about who will own and operate the transmission system and who will pay for grid expansion and improvement. This third obstacle has become important since the 1990s, as the electricity industry has been restructuring so that it is partly regulated and partly competitive. Opposition, approval process, and economic and structural uncertainty affect one another. Approval processes provide opportunities for people to organize and oppose. Opposition, including litigation, can delay approval and force modifications in transmission design and routes and the prospect of this increases uncertainties, deters investment, and provokes project cancellations.

2.1 What Utilities, Government, Consultants and Researchers Say

In 2000, 23 persons across the United States affiliated with electric utilities and related energy-producing and marketing firms, energy-related government agencies, industrial consumers, and environmental groups were interviewed about the need for and obstacles to the expansion of electricity transmission capacity by Eric Hirst, a consultant on electric industry restructuring, for his July 2000 report to the Edison Electric Institute (EEI), the trade association of the investor-owned electricity utilities. Hirst's interviewees mention the following as cases of transmission facilities not being built because they were opposed and/or did not meet requirements for approval.

- A 765 kV line proposed by American Electric Power (AEP) in Virginia and West Virginia (*AEP 765);

- Two lines [230 kV (Chisago) and 345 kV (Arrowhead-Weston)] between Minnesota and Wisconsin (*A-W 345);
- A 500 kV line between Georgia and Michigan;
- Expanded interfaces between Indiana and Michigan;
- Expanded interfaces between Pennsylvania–New Jersey–Maryland and New York and between New York and New England [notably, three projects led by Connecticut Light and Power (CL&P): *CL&P 345 & cable];
- Additional transmission between Wyoming and eastern Colorado;
- Additional transmission facilities needed to serve the growing loads of Boston, New York City, Long Island, San Francisco, and San Diego.

EEI official David Owens, in the July/August 2001 EEI Electric Perspectives, attributes the problems of building new transmission facilities to the many challenges in regulatory review and approval. However, he indicates that these challenges are importantly driven by the public sentiment against transmission expansion and the involvement of government agencies, courts, federal and tribal governments, and competing interest groups. The public sentiment against transmission expansion, he says, is represented by the phrases NIMBY (Not in My Backyard), NOPE (Not on Planet Earth), and BANANA (Build Absolutely Nothing Anywhere Near Anyone). In other words, permitting processes enable people to express what has become a generic opposition to siting new facilities. Owens also provides examples of transmission projects that have been delayed for many years because of opposition and process, mentioning some of the cases listed by Hirst: the project by Xcel Energy (Chisago-Apple River) to link transmission facilities in Minnesota and Wisconsin, proposed in 1996, and the project by AEP to link West Virginia and southwestern Virginia through a 115-mile, 765 kV system proposed in 1990. He also notes the rejection by the Connecticut Siting Council of a cable to run below Long Island Sound, linking Long Island to Connecticut. By summer 2003, the Chisago project had been cancelled, the AEP outcome was uncertain, and the Long Island cable projects were on hold. The AEP and Long Island projects are mentioned in news reports about the August 2003 Northeast Blackout as examples of how parochial interests block construction of interstate transmission. The AEP and Long Island cases are also mentioned in a May 2002 National Transmission Grid Study for the U.S.

Department of Energy (DOE) as examples of how needed interstate transmission projects are significantly delayed through siting processes. The study indicates that the Federal Energy Regulatory Agency (FERC) may need to expedite siting.

In October 2001, Alan Schiber, Chairman of the Ohio Public Utilities Commission, testified before the U.S. House Committee on Energy and Commerce on behalf of the National Association of Regulatory Commissioners about transmission siting, especially about efforts to increase the federal role in this. He disagreed with “the presumption in Washington that the reason transmission systems are not being expanded or managed sufficiently to meet transmission need and prevent congestion is that State eminent domain and siting authority is inadequate.” Instead, “a major impediment to siting energy infrastructure in general and electric transmission in particular is the great difficulty in getting public acceptance for these facilities.” No matter where siting responsibility falls, with state or with federal government, siting energy infrastructure is not easy or amenable to a “quick fix.” Schiber notes that people object to siting such infrastructure because they do not want to bear the impacts or the costs of the facilities (even though they do want the electricity).

Basin Electric Power Cooperative, which serves member systems in North Dakota, where it is based, and eight other western states, tells its customers in newsletters of its success in building needed transmission facilities, notably in 2003 a direct current system intertie near Rapid City, South Dakota. In interviews, its officers explain that public opposition is not an obstacle to transmission siting and that they successfully negotiate agreements with landowners (although Basin Electric newsletters do refer to the problems of the “not in my backyard” resistance to siting new transmission). Instead, they say, “A primary obstacle to the construction of transmission is the regulatory uncertainty associated with the deregulation effort and improper pricing methodologies for the high-voltage transmission system.” In their October 2000 newsletter, Basin Electric’s Vice President of Government Relations summarizes the consequences of the uncertainties by rhetorically asking: “Why would you build it if you don’t know that you can use it and everybody has the same rights to it?” Basin’s marketing manager asks, “Who should pay for it? Will there be a return on investment?” To obtain a return, Basin Electric calls for pricing in which all users pay their *pro rata* share of the average cost of the grid, a method known as postage stamp pricing. This differs from the currently dominant license plate pricing, which recovers costs for transmission facilities from

local ratepayers. Basin indicates that postage stamp pricing is opposed by parochial interests, presumably by states or localities with traditional low-cost electricity that fear that expansion of the grid will raise their costs by exporting their supply.

2.2 Opposition in the 1970s

Opposition to electricity transmission and recognition of its significance is not new. In the 1970s, electricity utilities began building high-voltage transmission systems to carry electricity long distances and people strongly resisted some of these projects. A notable example is that of the 1974–1979 resistance to the siting and building of a ± 400 kV direct current (DC) line across west-central Minnesota by the rural electric cooperatives United Power Association and Cooperative Power Association, collaborating as Cooperative United, or CU (*CU ± 400 kV DC). Farmers, rural town-folk, and urban allies fought the line in the meetings and hearings of the siting process and by litigating in the courts, demonstrating in the fields, confronting surveyors, officials, local sheriffs, and state police, sabotaging towers, facing arrest, dealing with federal agents, and more. Reporters for state and national media covered the events. Researchers studied and reported the case in journals, books, and films, with some researchers calling it an example of a grassroots social movement. The line was built, but the struggle against it forced changes in legislation about siting such facilities.

Another case of note was resistance to the siting of a 155-mile, 765 kV alternating current (AC) line by the Power Authority of the State of New York (PASNY), connecting its system with that of Hydro-Quebec, in Canada. This project (*PASNY 765), proposed in 1973 and argued through hearings for several years, was also protested militantly through some of the same actions as occurred in Minnesota. Minnesota and New York power line resisters learned from one another through telephone, mail, exchange of newsletters, and visits.

Opposition made it difficult to build electricity transmission systems from the 1970s on, in Minnesota, New York, and elsewhere. In its August 22, 1988 issue, *Newsweek* reported that there was a virtual moratorium in the United States on building new high-voltage power lines because of “community opposition and environmental worries.”

2.3 Opposition Reported Worldwide on the Web

Opposition is not limited to the United States, as power line opponents are pleased to announce. From

the late 1980s through 2000, people opposing siting by Ontario Hydro of a 500 kV AC transmission line through their neighborhood in Bridlewood, Ontario, Canada (*Bridlewood O-Hydro 500) provided an information service to tell their local colleagues and later others across the world about why and how transmission lines were being opposed almost everywhere. This service appeared first as newsletters and later as a Web site dedicated to spreading the word about struggles against power lines and other technologies that produce electromagnetic fields. The site briefly summarizes approximately 50 cases in which the siting or operation of electricity transmission facilities was opposed by community groups or permitting authorities. Many of these cases are reported for localities across Canada and the United States, but also reported are cases in the Philippines, Spain, Morocco, Australia, Italy, Venezuela, Greece, Mexico, Ireland, and Antigua. The summaries are brief and not sourced. One purpose of this list and others like it is to advertise that opposition to transmission lines is legitimate and common.

The Bridlewood site was discontinued in 2000, a few years after the Hydro-Quebec transmission line was approved. However, people in southeastern suburbs of St. Paul, Minnesota, opposing Xcel Energy’s project to upgrade a 115 kV power line in their neighborhood (*Xcel 115 upgrade), include the Bridlewood list of world power line struggles in their Web site. These protesters of Xcel have been in the news for contesting the upgrading project through 4 years of hearings and litigation. Their Web site is an important node in a worldwide network of transmission line opponents. One of these is the group Revolt, organized in North Yorkshire, United Kingdom, in 1991 to resist siting by the National Grid, a transmission only firm, of a 75 km, 400 kV AC link from Lackenby south to Shipton, near the city of York (*North Yorkshire 400). Their resistance apparently helped delay project approval until 2002.

Note that above-mentioned cases marked with an asterisk are discussed throughout this article to illustrate opponent ideas and organization. The designations, e.g., AEP 765, CU ± 400 kV DC, are the author’s for this article and not necessarily used by utilities or government.

2.4 Counteropposition

Opposition is the main observable and most significant public response to transmission projects; however, there are some instances in which people

have organized as apparent grassroots proponents of specific transmission projects. Some of these proponents are criticized by opponents as front groups for industry, or “Astroturf lobbyists.”

The 1970s CU $\pm$ 400 kV DC case provides an instance of this. Midway during the conflict, scores of farmers and other rural people across Minnesota came to the state capital in St. Paul for a Line up for Power/Pro-Power demonstration of support for the line. These were customers and employees of local rural electric cooperatives that were served by the two associations building the line. Their cooperatives organized the event, bused them to it, fed them, and prepared placards for them to carry with slogans proclaiming the need for more power and mocking the slogans of line opponents. Opponents confronted these proponents on the steps of the capital and in debates with legislators. Most proponents seemed uncomfortable in their role and unable or unwilling to argue for the line as ably or fervently as opponents argued against it. The cooperatives did not repeat this effort.

In the late 1990s, as many people and groups in West Virginia and Virginia were resisting the proposal by AEP to build a 115-mile-long 765 kV transmission line through their neighborhoods, others organized to promote this AEP proposal. Ron Nixon reports in the December 7, 1997 *Roanoke Times* that the promoters organized as the Coalition for Energy and Economic Revitalization (CEER). CEER’s leader, Bill Tanger (who also chairs the Roanoke Valley’s Sierra Club), called CEER “the largest grass-roots coalition in Virginia’s history.” Its aim, he says, is to prevent widespread electrical blackouts in southwestern Virginia and to foster economic development in Virginia and West Virginia. However, power line opponents claim that CEER people are paid by AEP and that they are not grassroots activists, but rather a front group for AEP and what public relations watchdogs term Astroturf lobbyists. The purpose of such lobbying is to make a planned program look like a spontaneous explosion of community support. Regulatory documents show that AEP paid more than \$500,000 to CEER through Tanger and another \$200,000 to the advertising agency that he runs. In contrast, the AEP line opponents spent nearly \$300,000 over 6 years waging their resistance against the line.

2.5 Windmill Advocates

There is another type of advocacy for transmission line construction that is not considered a front for industry, but is supported by those otherwise opposed to new electricity generation and transmis-

sion. This is advocacy for facilities to transmit electricity from windmills and other renewable or “green” sources, particularly if these facilities are owned by farmers, ranchers, communities, and small businesses rather than large utilities. Wind/alternative energy advocates in the United States and Europe have long complained that transmission facilities are not adequate to incorporate electricity from these alternatives into the grid. In Minnesota, North Dakota, and South Dakota, states with considerable wind resources, some of the same people and groups who have opposed transmission lines from conventional generation facilities have supported the expansion of transmission and generation when these include electricity from wind turbines. In January 2003, Minnesota regulators approved a proposal by Xcel Energy to build high-voltage transmission facilities in southwestern Minnesota under conditions that require it to speed up its purchase of wind-generated energy by several years. The routes for the line remained to be permitted. Michael Noble, who directs an organization promoting sustainable energy, Minnesotans for an Energy-Efficient Economy, says that unlike power lines from conventional sources, which are controversial, this power line system will be supported by environmentalists and local communities benefiting from the generation and sale of wind energy. One strong supporter is George Crocker, who helped lead urban allies of west-central Minnesota farmers in their 1970s fight against the CU line and helped lead people in Minnesota and Wisconsin in resisting the Chisago and Arrowhead-Weston lines in the 1990s and early 2000s. Crocker says these activities are consistent because all help farmers and other ordinary people to make a living and persuade industry to improve environmental quality.

In summary, opposition is the most significant and observable public response to the siting of electricity transmission lines and opposition combines with siting approval processes and financial and regulatory uncertainties to block or delay the construction of new transmission systems.

3. HOW DO TRANSMISSION LINE OPPONENTS ORGANIZE?

The siting of a transmission line is usually opposed by many groups, some organized specifically and locally, others participating as allies advocating related causes. Groups are composed of many

persons and leaders who vary in their commitment and persistence. They oppose transmission lines for many reasons and interact in multi-issued communication networks. They act as interest or pressure groups to influence political decisions more than they act as a movement to change society and culture, but they have the potential to promote such change.

3.1 Participation in Groups: Cadres and Peripherals

Some people oppose transmission facility siting as individuals, but most do so as participants in groups. Over time, the organization of opposition is often composed of many groups of varying types that interact in networks. Key groups are those organized among neighbors to resist the line and its impacts in their locality. These local groups may cooperate through umbrella organizations to contest the line over its whole length and in state or other centralized hearings or litigation.

Participants in the groups organized to fight the line vary in their depth of commitment and the time and effort that they devote to the fight. The result is that a local struggle against a transmission system is usually waged by a few core activists and a larger number of peripheral participants. Core activists represent the group consistently in public hearings and media interviews and constitute its cadre. They organize and direct the campaigns through which peripheral participants raise money, sign petitions, send letters, lobby officials, participate in hearings, demonstrate, and monitor surveying and building. Core activists also manage group newsletters and Web sites. The numbers of people participating peripherally oscillate widely. Hundreds come out for some events, notably important public hearings, but retire when their campaign seems either to have succeeded, for example, in forcing a rerouting of the line, or to have failed. However, hundreds again appear when they learn about new threats. Furthermore, some people on the periphery become more committed and join the core, even leading it when others tire.

3.2 Groups Divide and New Groups Form

People oppose transmission systems for many reasons or concerns and core groups may try to embrace the widest array of these. However, when people feel their concern is ignored, they may form a new group

to express it. Also, people form new groups when they differ over tactics or use a different attorney. Furthermore, people tend to form groups in each of the many localities potentially impacted by a transmission project and as routes are decided or changed, those who escape impact often drop out and those who are newly impacted launch their defense.

3.3 Groups Organize a Division of Labor

Core activists usually organize a division of labor, allocating roles according to the capabilities and experiences of their participants. Often they also organize a weak hierarchy using the institutional models familiar to them. Someone becomes chair or president, another becomes secretary, and another becomes treasurer. Furthermore, core activists also name their group, commonly in words that describe their concerns and whose initials form a colorful acronym. Groups establish officers and organizational names in part to facilitate their participation in official public hearings and litigation, to raise and account for money, and sometimes to give their participants the legal protection of incorporation.

3.4 Many Groups with Diverse Issues Collaborate in Networks

Groups organized specifically to fight the line are often joined in resistance by representatives of groups already established to advocate other causes deemed related, such as environmental protection, environmental justice, native rights, or green energy. Also, representatives of established civic and commercial associations and local governments sometimes support opposition efforts. These various parties exchange information about transmission and companion issues with persons and groups across nations and the world sharing these concerns. The Internet facilitates this communication. A group's Web site details its resistance efforts and also links to Web sites of other groups that oppose transmission and generation facilities, promote alternative energy, or warn about the effects of electric and magnetic fields on health.

The groups organized locally and voluntarily to oppose a transmission line project are often criticized as having but a single issue, stopping a line, but over time some group participants work on many issues related to the generation, transmission, and use of electricity and the impacts of electricity technology on health and welfare. Opponents embrace many

issues because they find that the best way to oppose a specific energy project is to propose alternatives to it and to debate energy policy and the culture of energy use. Another reason people expand issues is that during their campaigns against the line, they collaborated with other groups concerned with these other matters. For instance, people organized to oppose the Arrowhead-Weston transmission line in Minnesota and Wisconsin allied with an environmental justice group, Just Energy, and the Split Lake Cree First Nation in Manitoba, Canada, to protest the impact on the Cree of the hydroelectric dams likely to transmit electricity over the lines. Thus, the line protesters expand their issues to include environmental justice and native rights.

3.5 Pressure Groups or a Social Movement?

Environmental justice and native rights activism, like environmentalism, are generally understood to be social movements. In contrast, transmission opposition groups are not referred to as movements, but rather as interest groups or pressure groups, terms that refer to their efforts to influence specific, local political decisions. A movement can usefully be defined as a group of people who are organized for, ideologically motivated by, and committed to a purpose that implements some form of personal and/or social and cultural change, who are actively engaged in the recruitment of others, and whose influence is growing in opposition to the established order within which it originated or which has controlled it. Transmission line opponents become more like a movement to the extent that they not only try to influence decisions about the siting of a specific line, but try to change energy policy and the culture of energy use among their fellows and elsewhere. They become more like a movement when they recruit ever more people to participate with them in promoting these changes and when their local groups interact with others holding similar ideas across the world. This potential is not yet realized.

Transmission line opponents are sometimes labeled environmentalists, participants in the environmental movement. However, protecting the environment is only one of the reasons people give for opposing transmission lines. As noted below, line opponents often complain that their farm or community is sacrificed to a transmission route because environmentalists and environmental protection legislation have exempted parks, forests, wetlands, or even scenic highway corridors from inclusion in the route.

3.6 Example of Organization and Constituency

Rural England Versus Overhead Transmission Lines (Revolt) provides an example of organization and constituency. Revolt has been led by a small core group since its founding in 1991 to oppose a 400 kV transmission system proposed to be sited in North Yorkshire, United Kingdom, by the National Grid. The Revolt chair, M. J. O'Carroll, writes that the core has worked especially with approximately 1000 households and 100 landowners and with them has obtained 8000 written formal objections and two petitions of 12,000 and 14,000 signatures opposing the proposals. Revolt describes its participants as a widespread mixture of people, urban and rural, living inside and outside the region affected by the line, as well as overseas. They range in income from affluent, to middle income, to poor; some are large rural landowners, some live in small towns and suburbs, and some are residents of government housing. Revolt has also been joined by local government authorities of main and minor parties. Its participants oppose the lines for many reasons. Some from urban Middlesbrough support the campaign because they value the surrounding countryside for recreation; others have different reasons. In the countryside, landowners and other residents worry that the lines will lower their property values, but some residents of a prosperous suburb of Middlesbrough favor the new transmission system because it will be followed by the removal of an older 275 kV line from their neighborhood.

Even though the line was approved and energized by 2003, a few Revolt leaders continue to communicate ideas about transmission lines and other electricity and energy issues through their Web site. The site is linked to the sites of transmission line opponents around the world. Revolt has been an important node in the network of electricity transmission line opposition and contributes ideas about change, which inform a movement.

4. WHAT REASONS DO PEOPLE GIVE FOR THEIR OPPOSITION TO TRANSMISSION LINES?

People express many reasons for opposing transmission lines. What are these reasons and how do utilities and their consultants interpret these?

4.1 Understandable Feelings?

The director of transmission and distribution operations for the Tampa Electric Company (TECO) tells reporters for the August 18, 2003 *Tampa Tribune* “I understand how people feel. No one would want a (power line) pole in their back yard.” TECO officials say that they had to build the line and its 125 ft tall, 3 ft wide concrete poles through Egypt Lake because the neighborhood grew up around their substation and right of way. They affirm that they strive to be a good corporate citizen (by lessening the impacts of facility construction), but have an obligation to provide economical and reliable power. Described by the article as “repentant,” the officials say they are open to discussing how they can make the process of siting facilities work better.

4.2 Not in My Backyard?

One of the ways electricity utilities try to improve process and public relations is by having consultants explain why and how people oppose their projects. One such consultant, Anne Gunning, writes in a 2001 Edison Electric Institute pamphlet that whereas many project opponents may resist simply because they are opposed to anything in their backyard, increasingly people identify “real reasons” for their opposition. Gunning lists the popular acronyms that are used (by critics to explain resistance to their projects) and suggests that these acronyms and reasons have evolved. In the 1970s, there was NIMBY. In the 1980s, there was BANANA. In the 1990s, there was LULU (Locally Undesirable Land Use), NOPE, and CAVE (Citizen’s against Virtually Everything). In the 2000s, there is TARR (There Are Real Reasons). TARR, according to Gunning, is something new and important, which utilities must recognize. Dismissing opposition as NIMBY will not advance a project, she says. People believe that science, law, or some other “real” proof is on their side, even though they also often agree that new energy infrastructure is needed. Gunning writes that the concerns can be about money, environment, health, or quality of life. Although she advises utilities to recognize this, she also states that the people holding these concerns “will not change their minds based on evidence.” Furthermore, her presumption that it took people until the 2000s to base their opposition on real reasons rather than selfishness or obstructionism is incorrect. People have argued real reasons throughout the history of opposition and certainly since the 1970s.

4.3 Many Reasons Summarized

People express their many reasons for opposing transmission line siting in many forums. In three of these, Web sites, public siting hearings, and agreements between utilities and publics, they summarize their reasons.

4.3.1 *Save Our Unique Lands*

Save Our Unique Lands (SOUL) was formed in 1999 to oppose a 345 kV transmission line, 250 miles long, from the Arrowhead substation in Duluth, Minnesota, to Wasau, Wisconsin. The line is proposed by Allete (formerly Minnesota Power) and Wisconsin Public Service. These utilities named the project “Power Up Wisconsin,” reflecting their claim that this will ease a power shortage in Wisconsin and improve the reliability of the transmission system. SOUL opposes the line for many reasons, synthesizing these on its home page. Combining reasons of voltage, length, appearance, need, and fairness, SOUL refers to this as “a giant 345,000 volt ‘bulk transfer’ electric and fiber-optics line, proposed by two private for profit utilities, to slice a 250 mile permanent scar across the face of beautiful northern Wisconsin.” Adding concerns about environmental justice, equity, and property rights, it calls the line a huge “extension cord,” planned to cross the Ceded Territory of the Chippewa in order to transfer massive electric sales to points south and east and on through Chicago. It warns that the project’s “great negative impacts of health risks, property devaluation and ruination of the ‘north woods’ would be carried on the backs of the 11,500 private property owners and the tribes, as the companies use eminent domain to ‘take’ private lands.” There is “blood on the lines,” because the electricity sold on the lines comes from “two devastating sources.” One is coal plants in North Dakota, whose emissions add to global warming and deposit mercury in the northern lakes, contaminating fish consumed by women and children. Another source is the hydroelectric dams of Manitoba Hydro, “mega-dams,” which have flooded an area the size of Minnesota belonging to the Cree. According to SOUL, traditional Cree communities have experienced “inhuman suffering” because of this. And, according to SOUL, the power carried by the lines is not needed “to keep the lights on” in Wisconsin, but instead is to be used to power cities elsewhere as well as proposed copper–nickel–iron mining in northern Wisconsin, mining that SOUL and its allies also oppose.

The line siting authority, the Wisconsin Public Utilities Commission (WPUC), did not determine that SOUL's reasons were sufficient to stop the line project, but instead found that the line was needed and approved a specific route for it by October 2001. But in November 2002, the American Transmission Company, a transmission-only firm, which will build and operate the line, filed a petition to reopen the case because its new estimated cost for the project was \$420 million rather than the approved cost (of \$165 million). The WPUC will go to new hearings and opponents have this cost increase as a new reason to object.

4.3.2 *Revolt*

In the United Kingdom, the group Revolt (it does not capitalize its acronym) also effectively summarizes on its Web site some of its main reasons for opposing the National Grid's proposed 400 kV transmission line in North Yorkshire:

1. It is not needed for public electricity supply, only for profit (for the electricity industry).
2. It will ruin large areas of treasured countryside.
3. It will ruin some families by massively devaluing their homes, without compensation.
4. It will promote uneconomic bulk transport of power from north to south, wasting valuable energy in generation and transmission, and cause environmental damage.
5. This conflicts with new government energy policy favoring efficient local generation combining cooling and heating.
6. It will cost consumers dearly, in capital to build and in revenue for increasing operating costs both for the line and for knock-on grid developments throughout England.
7. The great majority of landowners refused to grant wayleaves (rights of way), despite outrageous inducements.
8. There is overwhelming public and all-party local government opposition.
9. People fear possible health risks from electromagnetic fields, which new evidence is reinforcing.

4.3.3 *Hearing Testimony Opposing AEP 765 kV*

In 1998 and 1999, public hearings were held in localities in Virginia impacted by two possible routes for a 765 kV line and related facilities proposed by a subsidiary of AEP. In his October 2000 report of the hearings, the hearing examiner describes the wit-

nesses, noting that over 500 witnesses testified. Some of these witnesses were members of a number of groups organized expressly to oppose the line in various localities. Some were members of local governments along the routes. Others represented established civic organizations, such as historic preservation committees. The examiner writes that the overwhelming majority of the public witnesses were opposed to any kind of transmission line in southwestern Virginia and he summarizes their reasons as follows: The line is viewed as a symbol of corporate greed at the expense of the cultural attachment of the people to their land and the scenic beauty of the region. The land is mostly pastoral—rolling fields and hills with mountain backdrops, and cattle and sheep farms with streams running through them. The line is described as a means of transporting electricity generated in the Midwest to sell in the Northeast, enriching corporate buccaneers and their stockholders. Risks to health and environment include contamination of ground and surface water from herbicides sprayed to keep rights-of-way cleared of vegetation, increased cancer caused by electromagnetic fields (EMF) radiating from the line, increased air and water pollution from the coal-fired generating plants providing power to be transported on the line, and noise from the line. Building the lines and towers will threaten caves and sensitive rock formations and old growth forests. Demand-side management and local generation provide alternatives to the line. The routes being considered are unfair in that they skirt National Forest land at the expense of the homes and land of ordinary people and in some places at the expense of poor people. The project will harm tourism and its benefits to local economy. Additional reasons were also given.

4.3.4 *Agreement between Hydro-Quebec and Farmers in Canada*

A summary of many of the impacts that transmission systems have on agriculture is found in the December 2000 statement of agreement between the Quebec farmers' association [Union des producteurs agricoles (UPA)] and the giant utility, Hydro-Quebec, in Canada. Hydro-Quebec explains that it is a government corporation that provides electricity primarily through hydroelectric plants often located more than 500 km from consumers. To reach consumers, it has constructed a vast power transmission system and its transmission lines run from east to west across many farms and woodlands. Hydro-Quebec and UPA agree that constructing such lines can have the following temporary and permanent impacts:

Temporary impacts during construction include:

- Impacts associated with the staking of rights-of-way;
- Reduced crop yields due to soil compaction;
- Alteration of underground or surface drainage systems;
- Alteration of irrigation systems;
- Damage to ditches;
- Broken fences, which can hinder livestock control;
- Noise produced by construction machinery, which can affect poultry and fur-bearing animals;
- Disruption of crop operations;
- Loss of time (during negotiations, for example);
- Loss of revenue (cash flow) while awaiting compensation;
- Impacts on areas or elements located outside the right-of-way, such as damage to farm roads, debris from tree felling, ruts and soil compaction, damaged trees, waste materials; and
- Construction debris and other waste materials.

Permanent impacts include:

- Loss of farmland or woodland;
- Loss of revenue, which could compromise the operation's profitability;
- Loss of time (e.g., time spent in negotiations or driving around the structure);
- Risk of farm machinery running into the structures;
- Creation of enclaves;
- Usage restrictions and other limitations associated with easements;
- Alteration of irrigation systems;
- Changes to crop operations;
- Impossibility or increased danger of using airplanes in agriculture.

The document also mentions that studies are being conducted to identify and analyze the biological impacts of electromagnetic fields on human and animal health.

4.4 Rules, Process, and Influence Are Also Reasons for Opposition

The impacts identified in Quebec are like those expressed by farmers in west-central Minnesota in the 1970s about the CU ± 400 kV DC line. The Minnesotans, however, complained that their concerns about these impacts were not heeded by the utilities, the siting officials, or the environmental and transportation agencies. Like some in southwestern

Virginia, the Minnesota line protesters complained that these agencies were constrained by environmental protection legislation to be more concerned about protecting parks, forests, wetlands, and highway corridors from power lines than protecting farmland. In other words, it was not only the decisions about the line that were wrong, it was also the process of making the decisions that was wrong. This is a common complaint in transmission line cases. Often opponents of transmission siting decisions also complain that even though they have presented many good reasons why a transmission project should be denied or fundamentally changed, ultimately the project is approved, perhaps with minor modifications. Government siting authorities say that they do heed the concerns of opponents. But they face a dilemma. On the one hand, people show that they want and need more electricity at an affordable cost. On the other hand, people do not want to be impacted by the facilities required to provide this electricity. They must reconcile the diverse interests and also manage the conflicts among these interests. Facility opponents may recognize the dilemma, but also suspect that the siting process favors energy companies, with their vast financial and political resources, and favors growth in electricity supply, since this accompanies desired economic growth.

In short, people present many reasons for opposing transmission lines. Some reasons concern the rules and processes used to make the decisions and many reasons concern the characteristics of the line, the need for it, and its impacts on environment but especially on human health and welfare. In the next section, this article explains in more detail how people argue against the line with respect to its need, its size, and the structure of the utility industry; its appearance and threat to their sense of well-being; and its threat to their health and safety.

5. NEED: THE FIRST DEFENSE AGAINST TRANSMISSION LINE SITING

Opponents of transmission lines quickly learn that the first defense against the line project is to declare that is not needed. If it is not needed to serve the public interest, then there is no reason for government to allow the electricity utility or transmission company to march the line over private property or community resources, using the power of eminent domain, and impacting environment, health, and welfare. Such logic is institutionalized in government

siting regulations. In the United States, state siting authorities decide whether an electricity-generating plant or transmission line is needed before, or as, they decide where to place the facility and whether the impacts on environment and people are acceptable or can be made so. The authorities usually decide this and award “certificates of need” and “environmental compatibility” through siting processes at the state and local levels. Processes usually include public participation, such as meetings in which the public is informed about the project and hearings in which the public critiques it. States vary in the structures and procedures that they use to make decisions about need and route, with some using one agency and others using multiple bodies.

5.1 Need Is Public

Need is defined as public. That is, siting authorities must determine that the facility serves the public interest rather than the interest of the utility. In the United States, it has been the states that approve and regulate facilities within their borders and the people whose interests that must be satisfied by a decision by state authorities are the residents of this state. Opponents often talk about this when they ask questions about transmission systems that cross state lines. In Connecticut, for example, opponents of a projected transmission system that they suspect to be for the electricity supply of New York reminded their siting council at a 2002 hearing that “the state legislature has given the council jurisdiction to objectively balance the statewide public need for adequate and reliable services at the lowest reasonable cost to consumers with the need to protect the environment and ecology of the state for the construction and operation of facilities under its jurisdiction.”

The utilities applying for a facility will also argue the public interest, claiming that they have the state-mandated requirement to meet the demand for electricity and that this demand is demonstrated by the evidence of public use. The president and chief operating officer of the Wisconsin utility proposing the Minnesota–Wisconsin Arrowhead–Weston line explains (in an EEI brochure about facility siting) that “customers expect electricity to be there when it’s needed, so we have to make sure the delivery system can fulfill their expectation.” How is expectation measured? Regulations tell the WPUC that the proposed facilities are to satisfy the *reasonable* needs of the public for an *adequate* supply of electricity (author’s italics). Permitting agencies also often require that the applicant demonstrate that demand

cannot be met more cost-effectively through other measures, such as more local generation, instead of long-distance transmission, or energy conservation and load management. Utilities, especially their economic and technical specialists, present detailed statistics and forecasts to argue their case before the permitting agency.

However, the utilities must also communicate need to the public, including those opposing or considering opposing the line. This public, particularly the opponents, will not only respond to claims of the utility, it will offer counterarguments, including counter-statistics to refute projections of use and counter-concepts about expectations and the meaning of reasonable and adequate. It was partly public opposition to electricity facilities and critique of need beginning in the 1970s that led legislators in Minnesota to require that utilities in the 2000s show that their calculations of need include consideration of conservation and other alternatives to a proposed facility.

5.2 Need Is Argued Objectively and Subjectively and on Three Fronts

Need is argued using numbers. Numbers, statistics, graphs, calculations, and projections seem central to debates about need. The Connecticut Siting Council, for instance, puts public need and benefit at the top of its list of factors, which can be objectively presented and argued. However, need is also argued subjectively. For instance, the meaning of reasonable and adequate is open to debate, much as is the difference between needs and wants. Power line opponents certainly are willing to debate these and other questions about need to further their efforts to stop the line. Opponents challenge project need because this is a major defense and because they can find much about need to challenge even though proponents present need as manifest and can interpret opposition as “NIMBY” selfishness. There are three ways in which people contest the need for a project. One is by contesting projections of local electricity demand and calculations of supply, arguing instead that the electricity will be exported for a profit. Another is by claiming that energy demand can be met with alternative technologies or reduced through conservation. People also reject the rule that once a certificate of need is granted, need can no longer be contested.

5.2.1 Contesting Supply, Demand, and Fairness

Opponents contest the calculations that the utilities and government regulators use to argue that

electricity demand is increasing and must be met with new supply. Opposition groups often include people with the knowledge and skills to do this well. At a mid-1970s public meeting in Minnesota on the CU proposal for a ± 400 kV DC line, one of the proposal opponents, a farmer with a doctorate in physics and previous employment in the aerospace industry, critiqued CU's extrapolations from the electricity consumption curves in west-central Minnesota. Speaking for the county group No Power Line, he argued that the curves showed a leveling of demand during the early 1970s so that power line construction could be postponed until the mid-1980s, when the technology of superconductivity would be advanced enough to put the line underground.

In 2002 hearings before the Connecticut Siting Council, Woodland Coalition opponents of the CL&P proposals to construct a 345 kV transmission system in southwestern Connecticut also contested the utility's projections. One said in written testimony: "If the definition of need is to be defined by the anticipated power consumption of southwest Connecticut, further supported by CL&P's own calculations of power consumption growth over the past 20 years, then I assert to this council that CL&P has fallen woefully short of proving its need for the amount of capacity expansion that it seeks." This critic goes on to argue that there is no need for an eightfold increase in generation capacity unless the council has expanded its definition of need to include connection to a regional and national power grid. It is this issue of transmission of electricity across jurisdictions that informs the most powerful challenge of need and its determination, namely, that the need being served is someone else's.

The Woodlands Coalition line opponents complain that the utilities will use the new facilities to sell electricity to Long Island and New York. Connecticut people have paid or will pay for the generating and transmitting facilities and will suffer the impacts of these on their environment, health, and quality of life. Long Islanders have not done their fair share of building energy facilities. One of the Connecticut line opponents used his expertise as a financial analyst to argue that the utilities want to build the transmission not to meet local need but to make a profit. He finds this particularly true of the unregulated energy trading company component of Northeast Utilities, the firm that will build and operate the transmission systems. He says that it is as wrong to give an unregulated and for-profit utility the rights of Eminent Domain as it would be to grant this right to Wal-Mart.

Similarly, opponents of other interstate transmission projects contest need on the grounds that these projects serve the financial interests of utilities. To this, they also add charges of unfairness. The utilities will provide electricity to those who did not pay for the facilities and, in addition, by exporting electricity, local prices will increase. Although utilities customarily argue that there is a need for more transmission, some, as noted for Basin Electric, also say that deregulation has made it unclear whether those firms and customers who pay for transmission benefit fairly from it.

In the 1970s, well before deregulation, the west-central Minnesota farmers and rural townspeople opposing the $\pm$ CU line also complained that the line was not needed locally but instead was being used to power "the Cities," of Minneapolis and St. Paul. Sometimes these rural Minnesotans pictured the people of the Cities, especially the suburbs, as using electricity wastefully, but they also recognized that many of these people were their relatives and tried to enlist their aid in fighting big industry and distant government. This theme of energy flowing away, to cities, is a continuing one, expressed in 2000, for instance, in the portrayal of the Arrowhead-Weston Line as a giant extension cord to Chicago. Even without deregulation, utilities and siting agencies would have difficulty persuading people that transmission through their communities served their local needs.

One main response of the utilities to such problems is to argue that transmission across states does meet local needs by contributing to overall system reliability. A number of Edison Electric Institute publications during the 1990s make this argument. But it is an argument that is as likely to be as resisted as are other arguments for facility need, especially if the utility industry is increasingly put on a free market basis. James Creighton, a consultant on public participation for electricity utilities, wonders what will happen if state regulatory bodies say that it is unnecessary to issue a certificate of need for transmission lines because the market can decide need. This will reduce regulatory costs and may seem to speed up decision-making, something that some political leaders advocate. However, Creighton warns, the public is not accustomed to letting the market decide transmission need. It will challenge need and either the state will return to the issue or a plethora of local government entities will act to manage these challenges. The alternative to this is that the federal government, perhaps through the FERC, will act to determine need anywhere in the

country and perhaps also to make other transmission facility siting decisions. This has been advocated by some. But, as has been noted, state authorities resist this federalization because it threatens their traditional rights and they know it would be resisted by the public. Transmission line siting processes could become not only a battleground for making decisions about the need for and siting of electricity facilities but about state's rights.

5.2.2 Conservation and Alternatives

Another argument against the need for a new facility is that energy demand can be met with alternative technologies or reduced through conservation.

In the 1970s CU $\pm$ 400 kV DC case in Minnesota, line opponents pointed to technologies such as underground lines using superconductivity; wireless broadcasting of electricity using the ideas of Nikolai Tesla; and local generation using windmills, biomass, photovoltaic cells, and low-head hydroelectric dams. The quest for alternatives encouraged the power line opponents to ally with advocates of "soft," renewable energy sources.

Opponents of the generation and transmission of electricity through conventional technologies continue in the early 2000s to call for alternatives and to work with proponents of the latter. Sometimes the demand for alternatives seems initially a device to contest the need for a conventional overhead transmission line and legitimate resistance to it. However, advocates for alternatives have used the threat of public opposition to conventional facilities to induce utilities to include increasing amounts of wind and conservation in their energy mix. Mention has been made of the way that wind advocates work with utilities in the United States to support the need for more transmission lines if these also carry electricity from wind turbines owned and operated by farmers, ranchers, and rural communities. Efforts are also well under way in the United Kingdom and Europe to advance the use of wind resources by developing better ways to transmit electricity from these sources. However, in the United Kingdom, leaders of the group opposing the 400 kV line in North Yorkshire write that because wind resources are intermittent, incorporating electricity from them into a reliable grid is not efficient. Utilities must build too much infrastructure to do this, including more transmission lines. This argument is consistent with Revolt's criticism that large-scale transmission systems with remote power generation are inefficient. They lose vast amounts of energy transmitting over long distances. The need for new lines can be

diminished by reducing this inefficiency and improving the efficiency of energy use. Energy supply can be improved by using new technologies that combine energy functions. Wind should be used to make a locally storable form of energy, such as hydrogen by electrolysis of water.

5.2.3 Contesting Rules about When Need Is Argued

In addition to contesting the need for a facility, its opponents also often argue that because of procedural rules they are unable to debate need at appropriate times. On the one hand, people criticize efforts to site facilities before need is demonstrated. The case of the CU $\pm$ 400 kV DC line in 1970s Minnesota illustrates this. Because new state legislation governing energy facility siting had not yet fully taken effect, the siting authorities held hearings on where to route the line before they could hold hearings to determine the need for it. Those opposing the line argued that this made the hearings not legitimate and continually tried to question need during the routing hearings.

However, even though need is usually determined before debates about routes and impacts take place, opponents often never agree that there is this need or that they had adequate opportunity to argue against it. For one, fewer people come to the hearings to decide need than to those to decide a specific route. People are notified that they might be affected by a facility to meet such need because they are in a corridor from which a route could be selected. However, most do not come to hearings on need, apparently because they are busy, the corridor is wide, the threat of a route over them is distant, and the subject of need is general. Those few who do attend speak of the problems they have in persuading their neighbors to join them.

On the other hand, the rule that need, once decided, can no longer be at issue does not stop people from arguing against it when they are defending themselves against the actual threat of a line overhead. Furthermore, the rule itself is questioned. In early 2003, the Minnesota Center for Environmental Advocacy (MCEA) praised a change in the rules governing energy facility siting for not allowing approval of the construction of a new power plant or line when a less environmentally damaging alternative exists. MCEA argues, however, that continued useful and fair consideration of the alternatives is jeopardized by the rule. It envisages a situation in which people do not participate in deliberations on need because they do not know that this concerns them. When people find that the

facility may be built next to them, they can no longer raise the question of need and no longer do this by identifying and arguing for alternatives, such as conservation or different technologies. MCEA will likely lobby the legislature to change the rules.

6. "SIZE" OF THE TRANSMISSION SYSTEM AS A REASON

People find reasons and organizational opportunities to oppose transmission systems in what some term the size of the system and others term design and configuration. Size includes the width and height of towers, the length of the line, the voltage carried, and the complexities of the system. People can protest towers and lines of any size as invasive, but they usually find more reasons to protest larger towers and higher voltages. Furthermore, longer lines are more likely to impact more people, increasing the numbers and diversifying the skills of those who start or join local groups. Siting processes usually vary according to the size of the voltage and length of the line, with official opportunities for public participation (and thus opposition) often increasing as voltage and length rise over a certain threshold. States and other political units vary in their rules about size thresholds and in their criteria and procedures for making decisions. Thus, as lines extend across jurisdictional borders, they increase the exposure of line builders to different rules and procedures. Examples of how opposition groups respond to size are presented below.

6.1 High Voltage, Remote Siting, Long Distances

In their 1977 study for the Minnesota Department of Health on the health and safety effects of high-voltage transmission lines, researchers attribute much of the public worry about these effects to the fact that by the 1960s and early 1970s the electricity utility industry was constructing or proposing transmission systems much higher in voltage and much longer in reach than the public had hitherto experienced. These were 500 and 765 kV AC lines, then designated extra-high-voltage alternating current, and ± 400 kV DC lines, designated high-voltage direct current. Physical efficiencies, new technologies, and perceived new needs and opportunities gave utilities incentives to use such systems. Demand for electricity was growing exponentially; however,

utilities found it increasingly difficult to build generating plants near the main load (use) centers because of public opposition and regulations to protect air and water quality. One answer was to build power plants in remote locations. One such place was the Great Plains, where there were resources of coal or lignite fuels to be mined, river water for turbine cooling, and air that could suffer some degradation within clean air rules. Another was in Manitoba or Quebec, Canada, where there were rivers to be dammed for hydroelectric generation. New technologies enabled firms to build higher voltage lines from generators to load centers, taking advantage of physical efficiencies. The capacity of a line to carry power varies approximately as the square of its operating voltage. Thus, doubling voltage provides four times the power transfer on the same right-of-way.

6.2 1970s: CU ± 400 kV DC

Responding to these conditions, two rural electric cooperatives based in Minnesota, Cooperative Power and United Power, collaborated through the Rural Electric Administration to plan and eventually build a system of generation and transmission that was termed innovative even by critics. The electricity is generated at a 1100 MW Coal Creek station along the Missouri River at Underwood, North Dakota, fueled by lignite coal mined in a giant strip mine next to the station. The alternating current produced by the generator is converted into direct current and transmitted on a ± 400 kV DC power line 430 miles eastward to Delano, Minnesota, just west of the Twin Cities of Minneapolis and St. Paul, where it is converted to alternating current for distribution to customers of the cooperatives, organized as CU. CU officials could take pride in this large size, long expanse, high voltage, and complexity. But these characteristics provoked and facilitated opposition.

The opponents criticized the technology and built organization along its likely routes. They said that when CU used the designation ± 400 kV DC, they were trying to hide the fact that this meant the voltage between the two pairs of wires of the system is 800,000 V, not 400,000 V. If the voltage difference between one of the pairs of wires and the ground is +400 kV and the voltage difference between the other pair of wires and the ground is -400 kV, the total, they say, is 800 kV. Only one other system in the United States, on the West Coast, has as high a voltage. CU's opponents said that they should not be made guinea pigs for new power line technology.

People organized opposition groups in different localities along the long corridor of the line in Minnesota and also shared information with a group in North Dakota that opposed the strip mine and power plant. These different groups contested the project locally, using particular approaches, but also cooperated. People also attacked the need for and fairness of the line by declaring that its 430 miles ends not in their lands, but in the suburbs of Minneapolis and St. Paul—the Cities. Yes, they need electricity and are willing to bear the impacts for such, but this is electricity for others. Some pictured these suburban and urban others as energy wasters.

6.3 1970s: PASNY 765 kV AC

In another case of remote siting and long-distance transmission in the early 1970s, PASNY proposed a 765 kV AC line to connect PASNY's system with that of the Canadian utility, Hydro-Quebec, rich in power from its dams. In testimony in April 1977 before a New York State Assembly committee reviewing protested siting decisions, a protester recalled that when she and her neighbors in upper New York heard about the proposed line through a series of meetings in 1973, 1974, and 1975 they were “unconcerned,” supposing it would be “another power line like the 230,000 volt line in our area.” But when she and others learned that it would be much higher voltage, they worried. Furthermore, when they looked at the long route of the line, they complained that it serves not them, in upper New York state, but New York City.

6.4 1990s and 2000s: A-W 345 kV, AEP 765 kV, CL&P 345 kV

Voltage and length continue to matter in the 1990s and 2000s. People organized as SOUL to fight the A-W 345 kV line in Minnesota and Wisconsin picture it as a giant extension cord for Chicago. Another example is found in the response to the AEP 765 kV line from West Virginia to Virginia. Persons interviewed by the *Roanoke Times* reporters R. Nixon and L. Caliri for their 1997–1998 series on the AEP project worry about the size of the line. The operator of an apple orchard notes that he and his workers already put up with a 138 kV line in the orchard and a 345 kV line nearby. But these are “nothing” compared to the 765 kV line. “Nobody’s going to want to work under all that line.” A vice president of the National Real Estate Appraisal Institute said that

studies in some Florida counties show that high-voltage lines lower values in and near the easement.

In Connecticut in the early 2000s, the Woodlands Coalition opposed three related line-building projects of CL&P. Testifying before the state siting authority, a coalition activist who holds a Ph.D. in physics and heads a technology firm says that a 345 kV line is qualitatively different from 115 kV and lower voltage lines. It can produce a corona discharge, which produces potentially dangerous ionizing radiation, a health risk. The chair of the Norwalk Shellfish Commission testified that a 345,000 V system is used for long-distance transmission across open spaces and would be unsuitable for developed southwestern Connecticut.

6.5 Xcel 115 kV, More Risk from EMF?

Not all transmission line opponents point to high voltage as the problem. In St. Paul, Minnesota, the Power Line Task Force is contesting a proposal by a Xcel Energy to upgrade from a single to a double circuit its 115 kV transmission line running through their neighborhoods. Its Web site explains how 115 kV lines can be as dangerous as high-voltage lines, of 345 kV and above. Overall risks do rise as voltage increases. However, risks from EMF are associated with the amount of current flowing through the line rather than voltage. To deliver a given amount of power, utilities must push more current through low-voltage lines than through high-voltage lines. The magnetic field under a 115 kV line is often greater than that under a 345 kV line. Also, high-voltage lines are customarily built on wider rights-of-way than low-voltage lines so people live closer to them than they do to high-voltage lines. Thus, the site declares, low-voltage transmission lines (up to 69 kV) generally endanger human health more than do high-voltage lines.

In summary, people find reasons and opportunities in the size of transmission systems to oppose their siting.

7. STRUCTURE AND RESTRUCTURING OF THE ELECTRICITY UTILITY INDUSTRY AS REASONS

The structure of the electricity utility industry and changes in this structure provide people with reasons to oppose transmission lines.

7.1 Deregulation

In the United States in the 1970s, protesters argued that the electricity utilities were becoming too large and too vertically integrated. The CU project in North Dakota and Minnesota is an example. With loans and urging from the federal Rural Electric Administration, two large rural electric cooperative associations combined with a coal company to mine and burn coal and to generate, transmit, and distribute electricity. However, in accord with changes in federal and state legislation, the U.S. electricity industry has, since the early 1990s, been changing its structure away from such vertical integration, regulated primarily at the state level, to one shaped by competitive markets. By 2002, 23 states and the District of Columbia had enacted restructuring legislation and most other states were initiating or considering such a change. Utilities are dividing into regulated and nonregulated entities and are separating generation from transmission and transmission from distribution. Firms whose sole function is to build and operate transmission systems are established. Regulation is giving way to competitive markets. Markets enable firms specializing in trading to buy electricity where it is cheapest and sell it anywhere—if they can transmit it. Such transmission requires that transmission systems at the state level be integrated into regional and national grids. This calls for the building of new transmission lines. These are opposed.

In the United Kingdom, the electricity industry says that their country is a model in the development of open energy markets. A privatization process begun in 1990, shaped by the Electricity Act of 1989, has created a competitive market with open access to transmission and distribution networks, particularly in England and Wales. The grid owners (National Grid for transmission and 12 regional distribution companies for distribution) are regulated monopolies, whereas generation and supply are privatized.

Opponents of specific transmission projects find reasons for opposition in these structural features and changes. Examples of these reasons are noted in this article. In essence, people argue that new transmission lines are not needed to serve the interests of those impacted by the transmission and generation project, or the general public, but rather to profit the few who own the transmission, generation, or trading systems. This idea underlies arguments about need, fairness, and environmental justice.

7.2 Expanding Federal Control

Another structural issue that affects opposition is the projected expansion of the role of the federal government in siting and permitting electricity facilities and designing the system of electricity marketing. As argued in the 2002 National Transmission Grid Study for the U.S. DOE, the U.S. transmission system is a national public good that must be modernized by building new facilities. The study does not directly attribute failure to build these to opposition, but instead it focuses on flaws in the approval process that, among other things, fails to obtain public acceptance and fails to overcome the competing interests of local authorities and state and federal agencies, for instance, in land use. It also finds fault in business uncertainties, such as those mentioned by Basin Electric. As a remedy, the DOE expects that the FERC can play a larger role in siting and making rules. One role could be to grant designated entities the right of eminent domain to acquire property for rights-of-way when states do not. There are many who oppose such FERC expansion. The objections of the National Association of Regulatory Commissions have been noted. Also, some groups resisting specific transmission projects object to the expansion of the FERC and to deregulation and seek to influence state and federal legislation accordingly.

8. SENTIMENTS ABOUT VISUAL IMPACTS AND LOSS OF CONTROL AS REASONS

Power line opponents describe the way that transmission lines and towers mar landscapes and even utility executives agree that these facilities are “ugly.” Opponents also feel the lines and towers symbolize their loss of control.

8.1 Visual Impacts in Rural England

On their Web site, Revolt writers warn that the visual impact of 229 pylons—many taller than York Minster or the Statue of Liberty—will be catastrophic. They write that the overhead line will run through countryside of outstanding beauty, skirting a national park and impinging on ancient woodland, historic buildings, and sites of special scientific interest. They include a poem that William Wordsworth wrote in 1802, praising the magnificence of the horizon that the giant pylons will deface.

In interesting contrast, the National Grid Web site, under the title “Overhead or Underground?” includes a 1932 poem by Stephen Spender that likens pylons to whips of anger, to nude, giant girls that have no secret, and to the quick perspective of the future. However, National Grid also asks: “Everything being equal, wouldn’t we all prefer electricity cables to be buried out of sight than straddling the countryside on big, ugly pylons?” Since everything is not equal, undergrounding has its own environmental, financial, and operational costs, so that it is usually not a viable option for high-voltage transmission lines. National Grid says that undergrounding may be the best option where overhead lines are not practical, as in densely populated areas. However, National Grid did agree to underground a small segment of the North Yorkshire 400 kV line because officials agreed with opponents that the pylons and line at this segment would jeopardize the views of some especially cherished landscape features of rural and parkland and that thousands of people enjoyed these views. Revolt’s chair writes that one reason that Revolt has attracted the participation of many city people not under the line is that they do not want the countryside marred.

Appreciation of rural landscapes is strong in English culture. This motivates groups organized specifically to protect the quality of life and environment in rural England to include in their programs opposition to overhead transmission lines. For instance, the Campaign to Protect Rural England and the Friends of the Lake District and Cumbria have campaigned vigorously over past decades for overhead wires to be undergrounded to preserve the English countryside, especially special landscape areas. They argue that this has become economically feasible.

8.2 Visual Impacts in Minnesota

It is not only in the English countryside that people object to the appearance of transmission lines and towers. In their 1970s protest of the CU line, rural Minnesotans sounded somewhat like rural Englishers in decrying the way that the towers and lines would ruin their landscape, a landscape that they appreciated for its open farm fields and rolling hills dotted with lakes and woods. Expressing the sentiments of many, one said, “You see, these wide open spaces are why we’re here. If you have these big old monsters, you suddenly have an encroachment on your space. And it would destroy our peace of mind. It would destroy the quality of our life.”

8.3 Lines and Towers as Symbols of Invasion and Loss of Control

Such words point to another reason that people oppose transmission lines. They represent the invasion and taking of their land by powerful others. In their 2000 environmental impact assessment for the Arrowhead-Weston line, analysts for the Wisconsin Public Utilities Commission recognize aesthetics and visual impacts as legitimate public concerns. They recognize that the aesthetic effects not only interfere with public enjoyment of landscapes, but also represent an invasion of landowner rights.

Under the category of “social concerns,” the analysts note that people may feel that they have lost control of their lives when the towers and line are built on their land. When people learn that the line may expose them involuntarily to electric and magnetic fields and that these fields may pose risks to their health, they may interpret this as a threat to their freedom to choose their risks. The comments of power line opponents agree with these assessments. Landowners can be compensated with money for taking their land, but this may not compensate for the intangible threats to their sense of well-being.

8.4 Political Leverage of Sentiments

Sentiments about the ugliness of transmission lines and loss of power and control are important forces motivating people to oppose transmission projects. Utility officials and government regulators may accord them some legitimacy. However, public participation specialist James Creighton, claims that risks to health and safety are more potent politically than ugliness and invasiveness. He intimates that protesters know this and base their objections on such risks.

The article next examines health concerns, notably those about electric and magnetic fields.

9. HEALTH, SAFETY, AND ELECTRIC AND MAGNETIC FIELDS AS REASONS

Opponents of overhead transmission lines often claim that these lines and supporting facilities put their health and safety at risk. Of the health risks, those of exposure to EMF have become the most argued, controversial, and politically significant. Electricity utilities and health departments define

EMF as invisible lines of force that surround any electrical device, such as a power line, electrical wiring, or an appliance. Electric fields are produced by voltage and easily shielded by objects. Magnetic fields are produced by current and these fields pass through most materials. Both fields weaken with increasing distance from the source. EMF has become an issue in transmission line siting because findings from research beginning in the early 1970s have suggested that magnetic fields may be associated with increased risks of cancer and perhaps other health problems. Opponents of transmission lines have learned about this, spread word about it across their networks, argued it in siting hearings, and helped to make the effects of EMF an issue in science and energy policy.

9.1 EMF Concerns in 1970s in Minnesota and New York

This story begins in the mid-1970s in Minnesota, where people opposed the proposed ± 400 kV DC line of CU. As part of their effort, these opponents identified others who were opposing transmission lines elsewhere and they also gathered information about transmission systems and their impacts. Several individuals in the group No Power Lines (NPL) read papers on long-term exposure to low-level electric and magnetic fields by Andrew Marino, a biophysicist and attorney, and R. Becker, a medical doctor, and an English translation of a paper in Russian that mentioned Soviet studies on railroad workers exposed to fields from a 500 kV substation. The Marino and Becker papers reported that rats exposed to the fields suffer adverse health effects. The Soviet studies suggested adverse effects on humans.

In 1975, NPL participated in public hearings held by the Minnesota Environmental Quality Council (EQC) to make decisions about the CU proposal. NPL introduced as evidence correspondence from Marino and the Soviet paper. EQC officials and the CU lawyer discounted the latter because it referred to a study not in evidence, a translation, and not the original paper of the original study. The EQC gave more weight to the analysis by an expert witness for CU, which found no significant biological effects from exposure to electric and magnetic fields. NPL participants were angered by this decision and by the way their evidence was dismissed and their abilities demeaned.

NPL and other groups continued their fight sufficiently to keep the EMF and other health and

safety issues alive even after EQC issued a construction permit for the CU line. In September 1976, the groups, organized as a regional coalition, petitioned the EQC to “modify, revoke, or suspend” the permit on the grounds that the line “would be dangerous to the public health and the environment.” They cited the direct testimony of Marino and Becker before the New York Public Service Commission, which, responding to its own power line protest, was conducting major hearings on the health and safety effects of two 765 kV transmission lines in New York State. The EQC Board denied this petition from CURE. However, the Board told the Minnesota Department of Health (MDOH) to conduct its own study of effects. Protesters said findings would arrive too late to stop the project. The MDOH study, a survey of the literature, was reported in October 1977, after construction was under way. Its authors concluded that adverse biological effects had not been demonstrated to exist either empirically or in the laboratory nor had they been explained by a reasonable hypothesis. They also declared it a logical impossibility to state conclusively that there are no such effects.

A main source of information for the MDOH researchers was the Common Record Hearings of the Public Service Commission (PSC) of New York State. In 1975–1976, PSC conducted an exhaustive inquiry through testimony and cross-examination of 26 scientists and engineers into the health and safety effects of two proposed 765 kV AC transmission lines in New York. One would connect the system of PASNY with that of Hydro-Quebec. Strong resistance to this line did much to persuade the PSC to conduct this inquiry.

At the hearings and in subsequent rebuttals of the PSC findings, Marino and Becker testified that biological effects might occur from exposure to high-voltage transmission line electric and magnetic fields and recommended that the 765 kV lines as designed not be built. However, PASNY presented four specialists in medicine, radiation biology, and biophysics who testified to the contrary. The PSC Board decided to authorize the line. However, it reserved the right later to require protective measures to protect the public health and safety from exposure to fields, including a wider right-of-way than proposed and reasonable restrictions on operating voltage and loading. In contrast to its Board, the PSC staff found that “the 765 kV line will probably cause biological effects in humans exposed to them on a chronic basis.” It recommended relocation of residences and more research. The PSC decision was

challenged and testimony rebutted. It was not until January 1977 that the hearing was closed. Construction began but resistance continued, sometimes through militant demonstrations in the fields. These decisions and events were reported in the media and Minnesota opponents of the CU line learned about them through their New York contacts.

The Minnesota opponents continued to resist the line for 4 years after the EQC granted the construction permit in 1976. They held mass rallies at construction sites and the state capital. They picketed and harassed surveyors and workers, threatening some with violence. They confronted local sheriffs and then state police. Someone shot through the windshield of a security guard's pickup truck. Over 120 opponents were arrested but the charges against most were dropped and only 4 were convicted. As the line was constructed and tested in 1978 and 1979, some opponents shot out thousands of insulators and, over many months, some toppled 14 towers by removing bolts. The opponents called this sabotage, likening it to resistance against an invader and claiming that it was necessary to protect them from the intolerable effects of the line on their health. These effects, they declared in letters to their legislators, the governor, and the state energy agency, included headaches, skin rashes, nosebleeds, respiratory problems, dizziness, and fatigue. In addition, some reported that their livestock were aborting fetuses and producing less milk. Officials held hearings at which people told of these effects. A Department of Health researcher referred to studies by University of California (Berkeley) researchers to raise the possibility that the effects were caused by a concentration of positive ions in the air in the vicinity of a ± 400 kV DC line. These might cause a secretion of serotonin, a neural hormone that can produce the symptoms mentioned by the farmers. Power line opponents and officials then argued over how best to investigate these health effects and make decisions about the future of the line. The opponents said that if the line were allowed to continue operating while its effects on their health were being studied, they would be human guinea pigs. Officials did not agree to stop operation, but they did survey residents along the line about their interpretations of health effects and they did hire the Berkeley researchers to prepare updated reports about transmission line effects. Meanwhile, CU and state officials asked the federal government to help them by federalizing the line under the Rural Electric Administration and to make attacks on it a federal offense, subject to FBI investigation. No one was arrested or charged for

toppling the towers. By 1981, the line was completed and overt resistance had ended; the ± 400 kV DC system has been operating since then and, CU says, serving Minnesotans well. The new health effects studies were inconclusive, proving neither that people suffered adverse effects under the line nor that they did not. By the early 2000s, the debate about these effects had diminished in Minnesota, although some former opponents continue to say that the effects are real.

9.2 EMF Concerns in Ontario, Canada, 1984–2000

In Canada, residents of Bridlewood, Kanata, Ontario, outside Ottawa, began opposing proposals by the electricity utility, Ontario Hydro, to site two 500 kV lines through their neighborhood in 1984, approximately 2 years before they learned about the health effects of electromagnetic radiation and referred to these in lobbying against the line. According to their Web-published "History of the Bridlewood Hydro Line Struggle," these power line opponents learned about EMF risks by reading a 1986 article on the health effects of electromagnetic radiation in the *Ottawa Citizen*. This news, as well as the failure of efforts to stop the line through local government proceedings, led them to mobilize against the line as the Bridlewood Residents Hydro Line Committee (BRHLC). This group tried to stop the line by communicating information about its costs and risks, lobbying Ontario government, and suing in the courts. They also picketed a visit of the Premier and blockaded Ontario Hydro construction vehicles. However, the lines were approved and built. Yet, BRHLC leaders say that they did have some victories—notably, the raised awareness of the EMF health and safety issue, locally, nationally, and internationally. They talked about the issue on national television news programs and spread word about it through handouts and, later, the Internet.

Two registered nurses in the group researched health effects through libraries. Two electric engineers in the group worked with Ontario Hydro technical officials to gather and assess information. Others established a community information network through Internet sites using the new National Freenet. There, people and groups exchanged information about transmission line issues, particularly about EMF. Eventually this was institutionalized in a Web site as the Bridlewood Electromagnetic Fields Information service. The service includes an

extensive bibliography on electromagnetic fields and health. Its hundreds of references include reports in major science journals and utility industry publications of research findings about the characteristics and effects of EMF and popular articles warning about the risks of exposure to EMF.

9.3 EMF Issues on the Internet

Although the BRHLC lost its fight against Ontario Hydro, a few participants continued to maintain the information service site and the bibliography. When the original administrator of the site resigned in February 2000, others maintained it. BRHLC also founded an open, unmoderated mailing list on Yahoo for the discussion of electromagnetic radiation. Participants around the world tell about the health effects of EMF from transmission lines and other sources, especially cell phone transmission masts; about their struggles against the siting of these facilities; and about technical details, such as regulations governing exposure to EMF. In 2000, the site's founder discontinued participation, but others, graduate engineers, carried on. They consider it important that "the EMF community" has this as a means of communication that belongs to it "collectively" and that is "open and unfiltered." Thus, even though a particular struggle against a transmission line ends in defeat, some participants continue to express their concerns about transmission line issues.

9.4 EMF and Transmission Line Opposition Networking, 1999–2003

The Bridlewood Electromagnetic Fields Information Service Web site also contains links to other groups opposing transmission lines, just as these groups link to Bridlewood. EMF effects provide one of the main concerns that these various groups can share. As mentioned, one of the linked groups is the Power Line Task Force in the southeastern suburbs of St. Paul, Minnesota, which from 1999 through at least 2003 challenged the project by Xcel Energy to upgrade its 115 kV transmission line running through their neighborhoods. The group leader, a financial planner with a doctorate in economics, established a Web site to publicize the group's fight, but also to spread word of EMF issues around the world. The group objects to the size of the towers of the Xcel project; however, its Web site chiefly attacks the effects of EMF. Xcel Energy insists that the upgrading will reduce exposure to these fields,

but the protesters disagree. They cite studies, particularly one by the California Department of Health in 2001, that they interpret to show that EMF can cause childhood leukemia, adult brain cancer, Lou Gehrig's disease, and miscarriage. Xcel Energy disagrees.

The powerlinefacts Web site links to the Web site of the group Revolt, just as Revolt links to it. The Web site of Revolt critiques the many impacts of a 400 kV transmission system proposed in 1991 by the National Grid for North Yorkshire and in operation in 2003. Revolt's Health Issues pages links to some 20 sites on EMF subjects. Some offer published scientific assessments of health effects, such as a much-quoted report released in 1999 by the U.S. National Institutes of Environmental Health Sciences and reports of the International EMF Project of the World Health Organization. Other links are to media reports of these assessments, new research findings, scholarly discourse about EMF effects, and risk management theory and policy.

9.5 Scientific Studies of EMF and How They Are Interpreted

Revolt, Powerlinefacts, the Bridlewood Bibliography, and other opponent publications refer to the reports of scientific studies about EMF and other transmission line health effects. Since the 1980s, there have been many such studies. Researchers from many disciplines, especially epidemiology, cellular biology, biostatistics, and risk assessment, have reported their findings in scholarly journals. Governments, health agencies, and research institutes at national, local, and international levels have convened panels of researchers to assess these findings. Utilities, their opponents, and siting authorities all refer to these studies to make their arguments or decisions.

Many of these studies were reviewed, summarized, and synthesized in 2002 by an Interagency Working Group in Minnesota and presented as a White Paper on Electric and Magnetic Field Policy and Mitigation Options to inform decision makers. This White Paper exemplifies many EMF studies. It distinguishes among findings from epidemiological studies, findings from laboratory studies, and the determination of a biological mechanism for how magnetic fields may cause effects, notably cancer (childhood leukemia and chronic lymphocytic leukemia in occupationally exposed adults). The paper concludes that epidemiological studies have mixed results. Some show no statistically significant association between exposure

to EMF and health effects. Some show a weak association. Newer laboratory studies fail to show such an association. Furthermore, the studies, with one notable exception, fail to establish a biological causal mechanism. Thus, most studies conclude that there is insufficient evidence to prove an association between EMF and health effects. However, many studies also conclude that there is insufficient evidence to prove that EMF exposure is safe.

These conclusions lend themselves to diverse interpretation and each major party in transmission line siting cases imposes its own meaning on them according to its interests and understandings. In essence, opponents say that these research studies and assessments show that the risks of EMF from power lines are significant enough to make government regulators deny approval of projects that site overhead transmission lines near human habitation or at least delay approval until more studies are conducted to determine risk. A weak association still means that there is risk and it is unjust to impose this risk involuntarily on people. Some also say that the applicant utility must prove that the transmission lines will not put people's health at risk from EMF before a project is given approval, rather than require that the opponents prove that there is a risk. Utilities say that the reports demonstrate that the risks from EMF are not sufficient to deny approval. They also disagree that the burden of proof of no risk be on them. Government siting authorities usually say that the reports show that EMF risks are not sufficient to deny approval. Some also recognized that if the burden of proof of no harm is shifted to the applicant, this could establish a new principle in regulatory policy, the precautionary principle. However, the siting authorities sometimes indicated that exposures would be mitigated if later scientific studies showed this was necessary. Siting authorities, legislatures, and international agencies continue to request such studies.

Meanwhile, scientists continue their studies, but their contributions seem likely to give the parties to the issue more to argue about. Findings and analysis reported by researchers at Bristol University in the United Kingdom provide an example. It will be recalled that scientists could not determine a biomechanical explanation of how EMF might produce health effects. However, on the basis of their research in the 1990s, including 2000 experiments near pylons in 1999, Denis Henshaw and colleagues at Bristol hypothesized a biomechanism. They report that high-voltage AC transmission lines produce corona ions that charge particles (aerosols)

from automobile exhaust and other pollutants in the atmosphere, near and downwind of the pylon. The deposition of these particles on the skin and in the lungs is increased, elevating the risk of diseases such as skin and lung cancer. Henshaw's propositions have been communicated to the public through the mass media. Revolt and other opponents of transmission lines mention the "Henshaw hypothesis" on their Web sites and the hypothesis has been used by opponents to bolster their testimony against line siting.

The media reports that scientists working for the electricity industry refuted the Henshaw results. The industry also notes in public announcements that the U.K. National Radiological Protection Board found an earlier small-scale study of the same phenomena by Henshaw to be flawed in method and theory and its results "implausible and highly speculative." In addition, industry and some government regulators point to the results of other studies that offer epidemiological and laboratory evidence that they say contradicts the implications of Henshaw, including studies for the National Radiological Protection Board (NRPB) by the renowned Sir Richard Doll.

In 2003 in the United States, researchers at the Electric Research Power Institute (EPRI) requested funds to assess the Henshaw Hypothesis. EPRI is the research arm of the electricity utilities industry. The debate about the hypothesis and the effects of EMF exposure is likely to continue. For example, Revolt reports that Henshaw's research provides further evidence that the hormone melatonin, involved in the body's circadian rhythms as well as in certain types of cancer, is also affected by magnetic fields. Revolt says uncertainties about the effects of EMF call for the use of the precautionary principle in siting transmission lines. Referring to positions taken by the European Court of Justice, Revolt says precaution means that in the face of such uncertainty, preventive measures may be taken without having to wait until the reality of the risks become apparent.

9.6 Perception and Communication of EMF Risks

Some analysts argue that the public is becoming unduly alarmed about the risks of exposure to EMF because they do not understand the meaning of statistical associations and probabilities as reported erroneously in the news. In an online review in the *British Medical Journal*, Simon Chapman, Professor of Public Health at University of Sydney, Australia,

discusses responses to news of a 2001 NRPB study on EMF risk. This study estimated that additional risk from power lines meant an extra case of childhood leukemia every 2 years: from approximately 1 in 20,000 to 1 in 10,000. Chapman writes that this is a very small risk, but that the media in Australia and the United Kingdom reported it under alarming headlines such as "Power lines double cancer risk." Chapman writes that because of such exaggerations of EMF effects, anxious Australians living near power lines have called cancer agencies asking whether they should sell their house or have their children tested. Others, such as John Farley in his Quackwatch Web site, attribute public fears about EMF to the exaggerations of quacks, who sell devices to detect EMF or medicines to cure it, or to the alarmist writings of journalist Paul Brodeur, who claims that government and industry are conspiring to hide the truth about EMF risks.

The American Physical Society (APS) also finds the risks of EMF effects exaggerated and not understood. In 1995, APS, the major professional association of physicists in the United States, published its official assessment of the effects of EMF on health. It stated that unsubstantiated claims that EMF are threats to health provoke fears among people and these fears lead to expensive mitigation efforts and, in some cases, to lengthy and divisive court proceedings.

As noted above, health agencies at the state, national, and international levels seek to communicate what is known about EMF effects in order to facilitate decision-making. However, D. G. MacGregor, P. Slovic, and M. G. Morgan suggest from their psychometric research on perception and communication of risks that even careful and scientific presentation of the uncertainties about EMF exposure will increase people's concerns and that people interpret new evidence according to naive conceptions about how EMF interferes with normal biology.

9.7 Is EMF Fear a Cause or an Effect of Opposition to Transmission Siting?

The Chapman and APS statements indicate that it is fear of EMF that causes people to oppose transmission lines. It will be recalled that utility consultants say that people oppose transmission lines because they are ugly or simply because they do not want them in their "backyard," but express concern about EMF and other health effects as their reason because

health is politically more powerful and legitimate. Which of these suggestions about cause is correct? Probably neither. It is difficult to determine cause-and-effect relationships in complex social phenomena (just as it is difficult to determine these in complex biological phenomena). This article shows how concern about the effects of EMF is part of an array of concerns held and learned by transmission line opponents. It is impossible to know how much it is concern about EMF that induces people to oppose transmission siting or how much it is concern about other impacts that induces people to learn about EMF risks and present these as reasons for opposition. When people work with others to oppose transmission line siting, and do so in contest with proponents of the line, they engage in a process of cultural learning. They learn why and how to oppose. They develop a system of reasons and other ideas about why the transmission project is wrong and as they do they also develop an organization to express their ideas and fight to stop or alter the project. Some of them continue to work on EMF and other electricity and energy issues after their fight ends. It usually ends because the line project is approved. But the project is usually approved only after their resistance has long delayed it and altered it.

10. SUMMARY

Organized opposition, informed by key ideas or reasons, has been the most observable and significant response to efforts to site and construct electricity transmission lines in the United States, the United Kingdom, and Canada. This is probably also the case elsewhere in Western industrial democracies, that is, where authorities permit people to oppose projects important to industry and government. Opposition is one of three major obstacles to the siting and constructing of new transmission lines. Opposition affects and is affected by the other two obstacles, namely, the complexities of official processes of obtaining approval and the uncertainties of obtaining sufficient return on investment. People organize their opposition to a transmission project by forming a local group that interacts in communication networks with other groups opposing the project and with groups elsewhere holding related concerns. People develop many reasons for opposing the line and communicate these in many forums, especially in public hearings, Web sites, and newsletters. Their main reasons are these: The line is not needed

because there are alternatives to it in generation, transmission, and consumption. The line is too high in voltage and carries electricity to other states or regions. It will be used chiefly to profit its owners rather than to serve the public it impacts. It is ugly and invasive and represents a loss of control. Unfairness and injustice mark the way the line was approved and in how its impacts and the impacts of generating its sources of electricity are distributed. Building and operating the line damages agriculture, farm economy, and the environment. The line threatens the health and safety of people living along it, especially from its electric and magnetic fields. People develop these reasons into a system of related ideas as they argue with utilities and siting authorities. They are usually understood to act as interest or pressure groups. However, these groups might become parts of a movement of change as they work to change the culture of energy use as well as energy policy.

SEE ALSO THE FOLLOWING ARTICLES

Electromagnetic Fields, Health Impacts of • Environmental Injustices of Energy Facilities • Public Reaction to Energy, Overview • Public Reaction to Nuclear Power Siting and Disposal • Public Reaction to Offshore Oil • Public Reaction to Renewable Energy Sources and Systems

Further Reading

- Caliri, L., Nixon, R., *et al.* (1997/1998). A Question of Power, The Roanoke Times, <http://www.newrivervalley.com>. [Series on the AEP 765 kV Case]
- Casper, B. M., and Wellstone, P. D. (1981). "Powerline: The First Battle of America's Energy War." University of Massachusetts Press, Amherst, MA. [The CU $\pm$ 400 kV Case].
- Gerlach, L. P. Energy wars and social change. In "Predicting Sociocultural Change" (S. Abbot and J. van Willigan, Eds.), Southern Anthropological Society Proceedings No. 13. University of Georgia Press, Athens, GA. [The CU $\pm$ 400 kV Case]
- Gerlach, L. P. (1981). Protest movements and the construction of risk. In "The Social and Cultural Construction of Risk" (B. Johnson and V. Covello, Eds.), pp. 103–145. D. Reidel, Dordrecht, the Netherlands/Boston, MA.
- Gerlach, L. P. (1999). The structure of social movements: Environmental activism and its opponents. In "Waves of Protest: Social Movements since the Sixties" (J. Freeman and V. Johnson, Eds.), pp. 85–97. Rowan and Littlefield, New York.
- MacGregor, D. G., Slovic, P., and Morgan, M. G. (1994). Perception of risks from electromagnetic fields: A psychometric evaluation of a risk-communication approach. *Risk Anal.* **14**, 815–828.
- Minnesota State Interagency Working Group on EMF Issues. (2002). "A White Paper on Electric and Magnetic Field (EMF) Policy and Mitigation Options." <http://www.health.state.mn.us/divs/eh/radiation/emf/>.

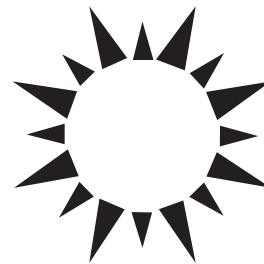
Opposition Group Web Sites

- Bridlewood Residents Hydro Line Committee (BRHLC): <http://www.feb.se/Bridlewood/>
- Power Line Task Force: <http://www.powerlinefacts.com>
- Rural England Versus Overhead Transmission Lines (Revolt): <http://www.revolt.co.uk>
- Save Our Unique Lands (SOUL): <http://www.wakeupwisconsin.com>
- Woodlands Coalition: <http://www.woodlandscoalition.com>

Public Reaction to Energy, Overview

ERIC R. A. N. SMITH

University of California, Santa Barbara
Santa Barbara, California, United States



1. Definition and Importance of Public Opinion on Energy
2. The Public's Knowledge about Energy
3. The Public's Environmentalist but Conflicted Tendencies
4. Public Opinion on Energy and Demographic Characteristics
5. Public Opinion and Political Orientations

Glossary

economic issues Political issues that relate to the distribution of wealth in society.

party identification A way in which people think of themselves; a psychological and emotional attachment to a political party.

social issues Political issues that relate to values and morals.

The term public opinion refers to the factual beliefs and preferences about issues of interest to the general population. Whether people believe that oil exists in the Arctic National Wildlife Refuge (a factual question) and whether people believe that the refuge should be opened to oil exploration (a preference) are typical examples of public opinion questions. Public opinion is normally measured using survey methods. Pollsters contact randomly selected samples of adults and ask them a series of questions to measure their opinions on current policy issues. The results of these surveys often play a major role in policy making.

1. DEFINITION AND IMPORTANCE OF PUBLIC OPINION ON ENERGY

What the public thinks about energy is important because public opinion is a major force influencing public policy on energy production and consumption.

The fact that elected officials generally follow their constituents' wishes on high-profile issues is well documented by political scientists. Studies certainly show the influence of major industries, such as petroleum and nuclear power, but that influence generally occurs on low-profile issues. Most studies show that on major issues that receive a good deal of press attention, such as nuclear power and offshore oil development, the public generally gets what it wants.

Academic studies aside, observers can easily see the influence of public opinion in the cases of both nuclear power and offshore oil development. Public opposition to nuclear power, especially after the 1979 Three Mile Island accident, helped block the construction of new nuclear power plants. In the aftermath of the accident, public support for nuclear power dropped sharply, and every proposal for a new nuclear facility was met with mass protests. Public opposition, coupled with soaring construction costs, ended the growth of the nuclear power industry. No more nuclear power plants were ever ordered.

In the case of offshore oil drilling, the California public has moved through intermittent periods of quiet acceptance and intense opposition during the past century. The 1969 spill in the Santa Barbara Channel crystallized public opposition, and—some historians argue—helped launch the modern environmental movement. During the 1980s and 1990s, public opinion polls consistently showed overwhelming opposition to further drilling. In order to curry favor with California voters, President George H. W. Bush ordered a moratorium on new oil drilling leases in federal waters in 1990. President Clinton continued it in his administration, and President George W. Bush has allowed the policy to stand. In the case of offshore drilling along the Florida coast, public opinion against oil drilling was so intense that in June 2002 President Bush ordered the Interior

Department to buy back some offshore oil leases that had been sold to oil companies and asked Congress to pass legislation to repurchase other leases. The official explanation for the buy-back from Department of Interior Secretary Gail Norton was that offshore oil development in Florida was opposed by “public opinion.”

One can also see the influence of public opinion on policy makers in every energy crisis from that of 1973–1974 to the gasoline and electricity price spikes of 2000–2001. Every time shortages appeared and prices soared, the public demanded laws to lower energy prices. Politicians leapt to do the public’s bidding, despite the fact that lowering energy prices encourages consumption, making the energy situation worse.

The initiative process gives voters a measure of direct control over public policy. In the 23 states and hundreds of counties and cities with the initiative process, voters can do more than demand that their elected officials pass the laws they want: Voters can pass the laws themselves. In these cases, public opinion controls energy policy. In California, for example, voters have passed both local and statewide initiatives limiting oil development. Although anti-nuclear power groups were not able to persuade California voters to pass an initiative to block nuclear power in the state in 1976, they came close enough to scare the state legislature into passing tough nuclear safety legislation that effectively blocked efforts to build more nuclear power plants. Moreover, the threat of an initiative vote against any controversial, proposed power source is clearly a deterrent to potential investors.

What the public wants does not always determine public policy. The influence of public opinion varies depending on a number of factors, including the attention given to the policy by the news media and the public, how much the public cares about the policy, expert opinion, and the interest groups arrayed for and against various policy choices. Nevertheless, in some circumstances, the public can get exactly what it wants—even against seemingly overwhelming opposition from the oil and gas, nuclear power, or electricity industries and their political allies. The public’s policy preferences may not always be wise or well informed, but they can be influential.

2. THE PUBLIC’S KNOWLEDGE ABOUT ENERGY

One of the best known findings of public opinion research is that most of the public pays little attention

to politics and, consequently, few people know much about politics or public policy. The list of facts not known by a majority of the public is stunning. What do the words “conservative” and “liberal” mean? What is the federal minimum wage? What majority is needed in the House and Senate to override a presidential veto? Who is the Speaker of the House of Representatives? Who is the Chief Justice of the U.S. Supreme Court? In national, representative surveys of adults, fewer than half the respondents offered correct answers to any of these questions.

These findings should not be exaggerated. The public is not ignorant or stupid. In fact, only a small percentage of the population knows virtually nothing about politics. Rather, most people do know something about politics, just not a great deal. The public’s modest level of knowledge stems from the fact that few people are interested in politics. In areas that people do care about—their jobs, their communities, their hobbies, and their favorite movies or sports teams—people know a great deal. Politics and public policy just do not draw much attention or interest.

Given these findings, one should not expect the public to know much about energy policy. There are a few aspects of the energy situation that almost everyone should grasp—for example, that the United States imports oil from the Middle East, whether gas prices are rising or falling, wind mills can be used to generate electricity, or waste from nuclear power plants is radioactive and hazardous. However, most aspects of energy policy are complicated and difficult to understand without some specialized knowledge or training. How much would new oil drilling in the Arctic National Wildlife Refuge or along the Florida or California coasts add to national oil supplies? Can the United States ever become energy independent? What are the effects of deregulation of the wholesale electricity market on prices? What are the trade-offs between clean air and electricity production? How safe is nuclear power? What is global warming, what causes it, and what are the policy options for doing something about it? None of these questions can be answered without a fair amount of knowledge. Because most energy policy questions are complicated, we should not expect the general public to understand them.

Table I presents selected survey results illustrating the public’s lack of knowledge about energy matters. Questions 1 and 2 show that despite the Arab–Israeli wars, the gas lines, and the sharp gasoline price hikes, in the 1970s only approximately half the public realized that the United States had to import oil to meet its energy needs. The failure to realize this

TABLE I
Selected Knowledge Questions, U.S. Samples

Question	Year: % correct
1. Do you think the United States has to import oil to meet our needs, or do we produce as much oil as we need?	1977: 48 1979: 51 1980: 63 1991: 50
2. Using the card, at the present time, how much of the oil used in the United States would you say we have to import from other countries? (Five choices)	1977: 29 1978: 30 1981: 27
3. About what proportion of its oil would you say the United States imports from other countries? Would you say two-thirds or more, about half, about one-third, or less than one-third?	1984: 33 1990: 49
4. There is a debate about whether this country should become less dependent on oil imported from foreign countries. Please give your best estimate of which phrase comes closest to describing how much of our oil comes from foreign countries—about a fourth, about half, about three-fourths.	2002: 34
5. Here is a list of different companies. All of them have operations here in the United States. But, would you go down that list and for each one would you tell me whether, to your knowledge, the main headquarters is located here in the United States or located in a foreign country?...	1978: 78 1986: 76
5a. Exxon?	1978: 17
5b. Shell?	1986: 19
6. From what you've heard or read, do you think a nuclear power plant accident could cause an atomic explosion with a mushroom-shaped cloud like the one at Hiroshima.	1979: 33
7. To your knowledge, what percentage of the nation's electric power is currently supplied by nuclear power plants?	1979: 5 1986: 6
8. In 1979, about 20 years ago, there was a nuclear power plant breakdown near Harrisburg, Pennsylvania. Would you happen to remember the name of that nuclear plant, or not?	1999: 38

Sources. Items 1–3 and 5–7 are from Delli Carpini, M.X., and Keeter, S. (1996). "What Americans Know about Politics," Appendix 3; Yale Univ. Press, New Haven; CT. Item 4 is from the Associated Press poll, February 22–27, 2002 ($N = 1016$). Item 8 is from the Gallup poll, March 19–21, 1999 ($N = 1018$). Both Items 4 and 8 are reported by Polling Report.Com (<http://www.PollingReport.com/enviro.htm>). Note: Delli Carpini and Keeter do not identify the surveys or sample sizes for the surveys, but all surveys were conducted by major reputable firms and are archived by the Roper Center for Public Opinion Research.

fundamental fact about the U.S. energy situation during the 1970s may seem astonishing, but it is consistent with other findings about the level of the public's knowledge. Large blocks of information are missing, and only a relatively small number of people have much detailed knowledge about any public policy issue. The other results in Table I add to the picture of the public's limited knowledge about energy policy.

Questions 3 and 4 show that the public lacks knowledge about our dependence on foreign oil continues. Question 5 shows that despite the political rhetoric about foreign and multinational oil companies, only a small portion of the public can identify them.

The next question about the possibility that a nuclear power plant could explode in a mushroom-shaped cloud is particularly revealing. The question was asked soon after the Three Mile Island accident. Even though television, newspapers, and news magazines reported extensively on what might happen during the Three Mile Island nuclear power

plant accident, only one-third of the public knew that a "mushroom-shaped cloud" was not possible. The Jack Lemon/Michael Douglas/Jane Fonda movie, "The China Syndrome," which premiered just before the Three Mile Island accident and offered a clear description of what might happen in the event of a nuclear power plant accident, may have been a critical and box office success, but an educational success it was not.

Table II presents results from a survey of Californians. The first question in Table II asks how long the world's oil supply will last. This is a difficult question because the topic has not been the subject of many news reports since the 1970s and because experts disagree. It has been the subject of some news coverage, however. Coincidentally, it was the subject of a *Scientific American* article published only weeks before the survey. Despite this, half the respondents admitted they did not know the answer and declined to guess. The answers from the other half ranged from less than 5 to more than 1000

TABLE II
Selected Knowledge Questions, California Sample

Question	% correct
1. As you know, the amount of oil in the world is limited. Do you know roughly how many years it will be before experts believe the world will run out of oil? [Probe if necessary]: Just your best estimate.	23
2. Do you happen to know whether it is currently safer to transport oil using oil tankers or using oil pipelines? By safer, I mean the way which is least likely to result in a major oil spill.	58
3. How often do you think a typical offshore oil platform along the California coast is likely to have a major oil spill—once every 5 years, once every 10 years, once every 20 years, once every 40 years, or less frequently than this?	15
4. When a major oil spill occurs, how much threat does it pose to human life—a great deal, some, only a little, or no risk at all?	6

Source. California Offshore Oil Drilling and Energy Policy Survey, March 1998, conducted by the Field Institute ($N=810$).

years. A reasonable estimate is somewhere between 50 and 100 years. Taking any number in this range as correct, we see that only 23% of our sample could make a reasonable guess.

The next three questions in Table II differ from the previous questions because they not only measure factual knowledge but also the public's perceptions of risk. In the abstract, this difference may sound trivial, but in practice people who feel threatened respond by demanding that politicians do something about the risks. In the case of oil, those demands are to limit or end offshore oil development.

The first risk question asks respondents whether it is safer to transport oil by using oil tankers or pipelines. The answer, pipelines, should be relatively well-known because most of the world's catastrophic oil spills have been tanker disasters, including such well-known ships as the *Castillo de Bellver*, the *Amoco Cadiz*, the *Torrey Canyon*, and the *Exxon Valdez*. Because tanker spills have received so much news coverage, one might expect most of the public to know this answer. The survey, however, showed that only 58% of the California respondents correctly identified pipelines as the safest way to transport oil. That is only slightly higher than the 50% who would have gotten the answer correct if they had guessed based on coin tosses.

The second risk question asks respondents to estimate how frequently a typical offshore oil plat-

form will have a major oil spill. The public's answers to this question reveal seriously exaggerated fears. Twenty-three percent of the California public said that they expected a major oil spill every 5 years from a typical platform. An additional 25% said that they expected such a spill once every 10 years. Although the exact meaning of "major oil spill" is not stated in the question, by any definition major oil spills from a typical platform are far less frequent than once every 10 years. If we were to use 1000 barrels of oil as the standard for a major oil spill, we would have to conclude that the correct answer—given by only 15% of the sample—was "less than once every 40 years" because there were only 11 spills of that size from all U.S. offshore platforms from 1964 to 1992 (and none since 1980). Even using a far smaller number of barrels as the standard would still yield less than once every 40 years as the correct answer because of the huge number of offshore oil platforms in operation in U.S. waters (more than 3800 platforms in 1990). This exaggeration of risk, or lack of knowledge, presumably contributes to the public's opposition to further offshore oil drilling.

The answers to the next question about oil spills also help us understand people's fears. The question asks how much threat a major oil spill poses to human life. Again, we see that the public seriously exaggerates the threat posed by oil. Thirty-one percent of the statewide sample responded that a major oil spill would pose a serious threat to human health. An additional 36% believed that it would pose some threat. In fact, oil spills pose only a negligible threat to human life. They certainly pose a substantial threat to many other kinds of life but not to human beings.

The public's low level of knowledge about energy policy is important because of the public's influence on policy makers. When the public forms opinions and makes demands on politicians based on misinformation, public policy may suffer. As the data in this section show, the public is often misinformed about even the most basic aspects of energy policy.

3. THE PUBLIC'S ENVIRONMENTALIST BUT CONFLICTED TENDENCIES

When the public does turn its attention to energy issues, it tends to prefer environmentalist policies. That is, the public generally favors conservation;

favors alternative, sustainable energy sources (wind and solar power); and opposes conventional energy development (oil, nuclear power, and coal). However, the public also displays a clear preference for government action to maintain low energy prices, despite the fact that low prices lead to excessive and wasteful energy consumption. These inconsistent preferences seem to result from a mix of people's misunderstandings about the energy situation and their distrust of the energy industry.

Evidence for the environmentalist tendencies of the public can be seen in a variety of survey results. Figure 1 shows the percentages of people who said in response to a series of Gallup polls that they preferred more conservation or more production as a solution to U.S. energy problems. The solid line represents the percentage indicating that they preferred conservation, and the dotted line represents the percentage indicating that they preferred production. Even at the height of the gasoline price hikes in 2001, the public preferred conservation by a substantial margin. By March 2002, when prices had stabilized and the public was becoming accustomed to higher prices, conservation was preferred by a margin of 60 to 30%. (The remaining 10%, not shown in the figure, either responded "both" or "neither.") Figure 2 shows the results of a related question in which respondents were asked in March 2001 and March 2002 whether they would give priority to development of energy supplies or protection of the environment. The environment won by a 52 to 40% majority. Figure 3 presents another trade-off. Respondents were asked whether they believed "protection of the environment should be given priority, even at the risk of curbing

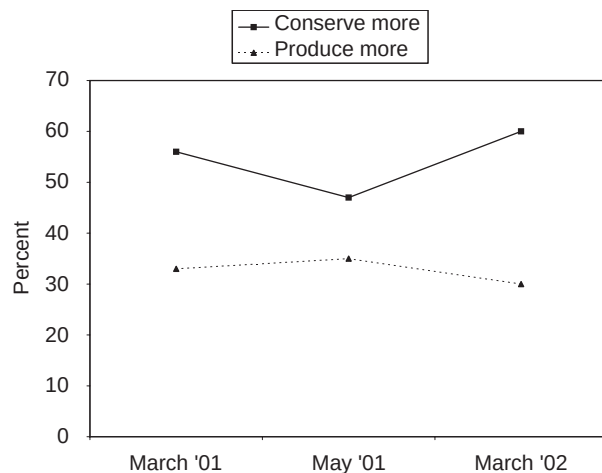


FIGURE 1 Energy solution: More conservation or more production. Source: Gallup polls.

economic growth, or economic growth should be given priority, even if the environment suffers to some extent." By wide and consistent margins from 1984 to 2000, the polls showed a strong public preference for environmental protection. Support for environmental protection ranged from 61 to 71%, whereas the number giving priority to economic growth ranged from only 19 to 32% and was only 28% when the question was last asked in April 2000.

On more specific policy questions, polls show that the public strongly supports investment in environmentally friendly energy sources and generally opposes the development of any conventional energy

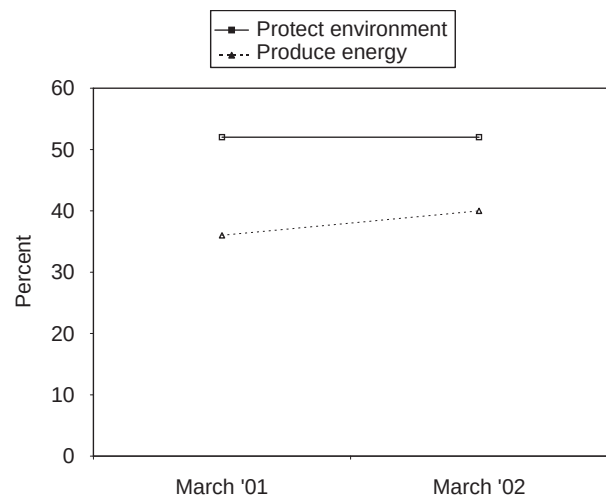


FIGURE 2 Environmental protection vs energy production. Source: Gallup polls.

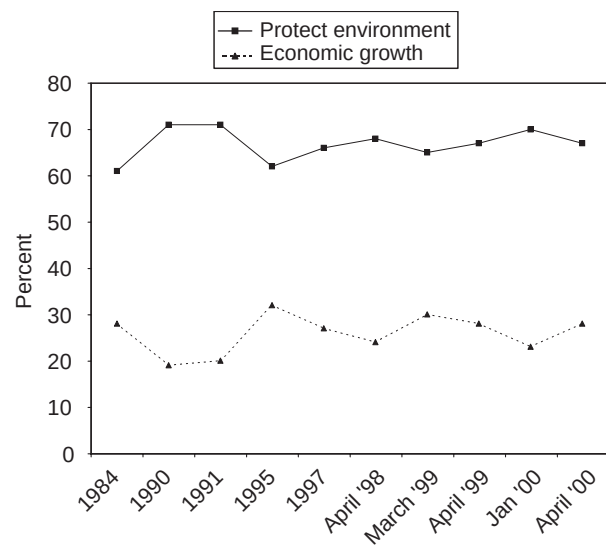


FIGURE 3 Environmental protection vs economic growth. Source: Gallup polls.

sources that are criticized by environmental groups. Table III shows the public's response to several policy proposals for dealing with U.S. energy problems. The public is nearly unanimous in calling for more support for solar, wind, and fuel-cell technology. Requiring automobile companies to build more fuel-efficient cars is also strongly supported. The proposals that fail to win majority backing are drilling for oil in the Arctic National Wildlife Refuge and nuclear power.

Further questions in other polls show majorities opposing drilling for oil and natural gas along the California coast or weakening environmental protections, such as automobile or industrial emissions, in order to increase energy production.

Despite evidence that a substantial majority of the public holds environmentalist views, there are signs that the public does not consistently follow through on those stated preferences. The public supports conservation in principle but opposes a variety of conservation policies because they place burdens on consumers. Simply stated, the public wants the benefits of cheap, plentiful energy but does not want to pay the costs.

The best behavioral evidence of the public's conflicted views can be seen in sales figures for sport utility vehicles (SUVs). According to a March 2000 Gallup poll, approximately one-fourth of U.S. households own at least one SUV. Moreover, in 2000—a year of sharply rising gasoline prices—nearly half of all new car sales were gas-guzzling SUVs, minivans, or light trucks. Thus, although the public says that it wants fuel-efficient cars, it actually buys fuel-inefficient cars (and trucks and SUVs).

Survey data also reveal the public's desire for low-cost energy. During the spring of 2000, when oil prices surged, a Gallup poll found that 80% of the

public favored lowering state fuel oil taxes, 75% favored reducing federal gasoline taxes, and 59% favored drawing oil from the U.S. Strategic Petroleum Reserve to put more oil on the market and cut prices. In a similar vein, when asked whether they favored or opposed the proposal of "giving tax breaks to provide incentives for drilling for more oil and gas in the U.S.," Gallup poll respondents in March 2001 favored the idea by a 53 to 43% margin. In addition, the proposal to set "legal limits on the amount of energy which average consumers can use" was solidly opposed by a 62 to 35% margin. The 2000–2001 electricity crisis in California provides further evidence of the public's desire for cheap energy. A May 2001 Field Institute poll asked Californians, "Is it a good thing or a bad thing for the federal government to cap energy prices?" Seventy percent said it was a good thing.

Most economists tell us that capping or subsidizing energy prices during shortages is foolish. If prices are allowed to rise freely, consumers will use less energy, and the energy problems will be resolved by free market forces. In contrast, if prices are capped or if energy consumption is subsidized, more energy will be used and shortages will worsen—potentially leading to lines at gasoline stations, electricity blackouts, or other problems.

Unlike economists, most of the public likes energy subsidies and dislikes free market solutions to energy shortages, but this does not mean that people are rejecting free market economics. Instead, many people believe that energy shortages are faked by energy companies to increase prices and profits. A 1974 Roper poll showed this when it asked, "Some people say there is a real shortage of gasoline and fuel oil because demand has outrun supply. Others say there really isn't a shortage of gasoline and fuel oil and the big companies are holding it back for their own advantage. What do you think—that there is or is not a real shortage of gasoline and oil?" Seventy-three percent said there was no real shortage. During the next energy crisis in March 1979, the CBS/*New York Times* poll asked, "Do you think the shortage of oil we hear about is real or are we just being told there are shortages so oil companies can charge higher prices?" Sixty-nine percent responded that they were just being told that shortages existed, and another 11% said they were not sure. Only 20% believed the shortages existed. Similar survey findings were reported during the gasoline price hikes immediately before the 1991 Persian Gulf War, during the 2000–2001 gasoline price hikes, and during the 2001 California electricity crisis.

TABLE III
Public Support for Energy Alternatives

Alternative	Favor (%)	Oppose (%)
1. Investments in new sources of energy such as solar, wind, and fuel cells	91	8
2. Mandating more energy-efficient cars	77	20
3. Opening up the Arctic National Wildlife Refuge in Alaska for oil exploration	44	51
4. Increasing the use of nuclear power as a major source of power	42	51

Source. Gallup poll, November 8–11, 2001.

Apparently, the public's reasoning is that the energy industries cause supply shortages and high prices by manipulating energy markets. In other words, high prices are the result of illegal or unethical activity. Given this situation, people can easily jump to the conclusion that the government should step in and fix the problem with price caps, subsidies, or other measures.

Here, the public's lack of knowledge about energy policy plays a role. In every energy crisis the United States has faced, politicians and activists have accused energy companies of manipulating prices. Certainly in the case of the oil crises, when people who do not even know that the United States needs to import oil to meet its energy needs hear those charges, they will be especially likely to believe them. More broadly, a mix of people's misunderstandings about the energy situation and their distrust of the energy industry may lead to poor judgments about the best energy policies for the government to adopt.

Public opposition to increases in energy prices affects more than government behavior during energy crises. Proposals for a carbon or Btu tax, such as the one offered by President Clinton in 1993, face a skeptical public. The idea behind Btu taxes is to increase the price of fossil fuels, causing consumption to decline and consumers to switch to using other fuel alternatives. Such proposals face two problems in winning public support. First, the proposals are complicated and require at least a moderate understanding of both economics and energy. Second, the proposals call for raising energy prices. These are both serious obstacles to Btu taxes.

In summary, the public generally favors environmentalist policies such as conservation and the development of alternative, sustainable power sources. However, the public also wants the government to take steps to guarantee low, stable energy prices. The contradictions among these views remain unresolved.

4. PUBLIC OPINION ON ENERGY AND DEMOGRAPHIC CHARACTERISTICS

To understand the sources of people's opinions about energy policy, it helps to examine the larger context of public opinion across a range of issues. Public opinion scholars have found that the best way to describe public opinion on domestic issues is to divide issues into two broad categories—economic

and social. Economic issues are those that relate to the distribution of wealth in society—the tax system, Social Security, Medicare, welfare, regulation of businesses and unions, etc. Social issues are those that relate to values and morals—abortion, birth control, free speech, civil rights, women's rights, sexual preferences, etc. This distinction is useful for two reasons. First, people tend to have similar opinions on the two types of issues. That is, if someone is liberal on one social issue, he or she is likely to be liberal on others. Second, the two types of issues have different causes and demographic correlates. Opinions on economic issues generally vary with income and self-interest, whereas opinions on social issues usually vary with education and age.

At first glance, energy policy would seem to be a classic economic issue. Energy production is a huge economic sector. Most disputes over energy policy involve the government regulating private businesses. Regulations affect company profits, create new jobs for some people, and eliminate existing jobs held by other people. From pollution controls to protections for endangered species, regulations on the energy industry have substantial economic impacts.

In some respects, attitudes toward energy policy do respond to events in ways that are consistent with people looking at energy policy as an economic policy. An excellent example of this is the way public support for offshore oil drilling along the California coast parallels the inflation-adjusted price of gasoline. Figure 4 shows a fairly close fit between support for offshore oil drilling and the real price of gasoline. When the price of gasoline increased in the late 1970s, so did support for offshore oil drilling. When the price of gasoline decreased in the 1980s, support for offshore oil declined as well. The 1989 survey shows a sharp decline in pro-development feelings that is not matched by gasoline prices, but it does

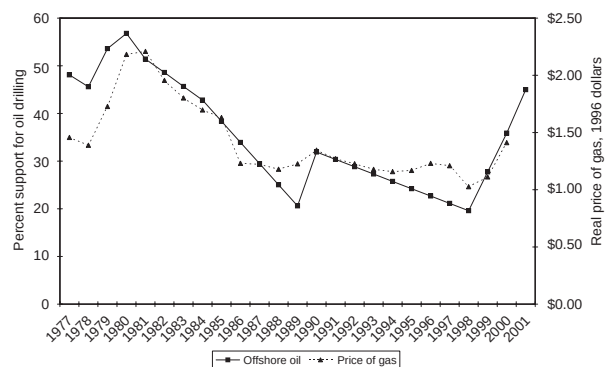


FIGURE 4 Support for oil drilling and the price of gasoline.

correspond to the Exxon Valdez oil spill. The post-1998 increase in support again parallels the increase in gasoline prices.

There is also some individual-level evidence that people think of energy issues in economic terms. For example, as the distance people commute to work increases, they become more likely to favor offshore oil drilling and other steps to increase oil production and cut prices. That is, economic self-interest affects attitudes toward energy policy in some cases.

However, the classic patterns of economic issues do not appear. Opinions on energy issues generally do not vary with income, occupation, or other measures of social class. A few studies have found that attitudes on energy policy vary with income, but these studies are outnumbered by the studies that found no relationships.

For the most part, attitudes toward environmental issues, including energy issues, behave like attitudes toward most social issues. Two key indicators of this pattern are the relationships that energy attitudes have with education and age. On virtually every social issue, the most liberal people in the population are the well educated and the young. The basis for the education–social issue relationship seems to be that education brings people into contact with new ideas and new types of people and broadens the perspectives of the well educated. This process develops a greater tolerance for people who are different—blacks, the disabled, gays and lesbians—and a greater appreciation of ideas such as the value of the environment.

Age is also related to opinions on social issues, but not because age causes beliefs or opinions. Instead, age is related to opinions on social issues because age reflects the generations in which people grew up and were socialized. Throughout the past century, the United States has been moving in a more socially liberal direction. Ideas such as racial and gender equality, which were once considered radical, are now generally accepted. Ideas such as gay rights, which were once never even discussed, are now on the leading edge of social change. Similarly, the modern environmental movement began in the 1960s. Therefore, people who grew up and were socialized before the 1960s are less likely to hold environmentalist views than those raised during the environmental era.

The relationship between education and the trade-off between protecting the environment or finding new sources of oil and gas is typical of relationships between education and questions about energy policy. Figure 5 presents the results of an April

2001 *Washington Post* poll, which showed that 51% of Americans with a high school education or less said that environmental protection was more important than finding new oil sources, whereas 58% of college graduates held that view. The better educated the respondents were, the more they favored protecting the environment.

The relationship between age and the environment–energy sources trade-off in the *Washington Post* poll is even stronger. As Fig. 6 shows, 66% of the 18- to 30-year-old respondents favored protecting the environment, whereas only 38% of those 61 or older did so—a 28% difference across the

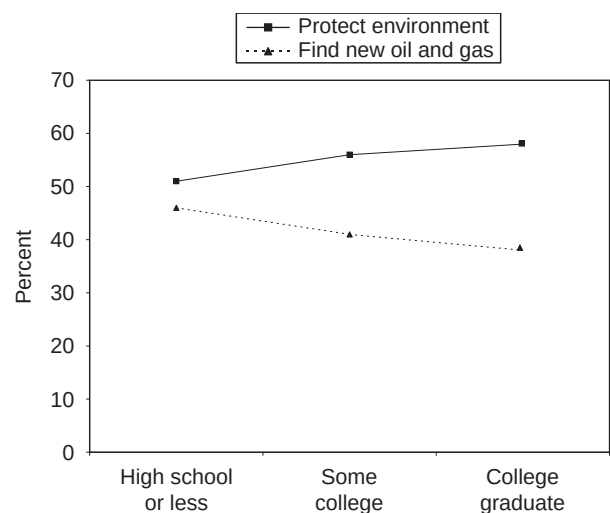


FIGURE 5 Energy vs environmental protection by education. Source: *Washington Post* poll, April 2001.

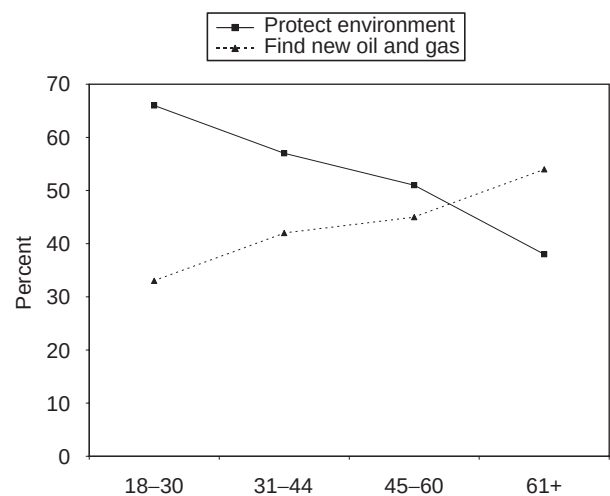


FIGURE 6 Energy vs environmental protection by age. Source: *Washington Post* poll, April 2001.

generations. This finding is typical in that age has been found to be the demographic characteristic yielding the strongest, most consistent differences on environmental issues.

Other demographic variables are also related to opinions on environmental and energy-related issues, although not as strongly or consistently as education and age. Gender is usually related to attitudes toward energy issues, but only if the issue entails some aspect of risk to people. When risk is involved, women tend to take the pro-environmental stand. For example, women tend to oppose nuclear power more than men do because many people in the public regard nuclear power as risky. In contrast, men and women usually hold similar opinions about offshore oil drilling because it is not seen as inherently risky by the public.

Race, ethnicity, and whether one lives in a rural or urban area have also been found to relate to environmental and energy issues in a number of studies, but the relationships do not consistently fall in a single direction. That is, in some studies blacks have been found to hold more pro-environmental views than whites; in other studies, the reverse pattern has been found. Studies of urban and rural dwellers yield similar results. The relationships often can be seen to depend on the exact issue in question. Location of hazardous waste dumps, for example, is of more concern to people living near proposed sites than those living at great distance. The demographic characteristics of those living near the dumps, therefore, may determine what studies find. Consequently, broad generalizations are difficult to make.

5. PUBLIC OPINION AND POLITICAL ORIENTATIONS

A second way to explain public opinion on energy issues is to examine people's political orientations and worldviews. By far the most important political orientation is party identification. Party identification is a way in which people think of themselves—a psychological and emotional attachment to a political party. People identify with political parties in the same way they identify with ethnicities or religions. Native-born Americans typically acquire a party identification from their parents by age 9 or 10—long before they understand anything about the substance of politics.

The importance of party identification stems from the way in which it influences people's opinions and behavior. Party identification influences people's opinions because it guides their decisions about what

to believe or disbelieve when they listen to the news or political discussions among friends. People who think of themselves as Democrats, for example, are far more likely than others to believe Democratic politicians and doubt Republican politicians. Similarly, Republicans trust fellow Republicans but not Democrats. As a result, both groups' view of reality is filtered through their partisan beliefs. In addition, party identification influences people's values. When a politician of one's political party says that preserving the Alaskan wilderness is vital, or that America desperately needs Alaskan oil, one's opinion is likely to be swayed. As a result of differing beliefs about the facts and differing opinions, the behavior of Democrats and Republicans differs as well. In voting and in a wide range of political activities, Democrats and Republicans normally choose opposing sides.

In the world of politics, the Republican Party generally represents business interests and the wealthier segments of society, whereas the Democratic Party represents the working-class and lower income segments. Because environmental regulations are constraints on businesses, Republicans have usually been far less supportive of them than have Democrats. As a result, Republicans have generally leaned in a pro-development direction, whereas Democrats have generally leaned in a pro-environment direction.

These partisan patterns can easily be seen both in national political disputes and in public opinion data on energy issues. The fight over whether to open the Arctic National Wildlife Refuge for oil exploration and drilling is a good example. In the 2000 presidential election, the issue that drew the most public attention was the rising price of gasoline. Other issues garnered more press coverage, but drivers were reminded of the high prices every time they filled their cars' gas tanks. As a result, public interest remained high throughout the year. Both presidential candidates responded to the public's concerns. Governor George W. Bush, a former Texas oil man, declared that the way to lower prices was to produce more and, in particular, to open the Arctic National Wildlife Refuge in Alaska for oil drilling. Vice President Al Gore countered that the way to cut gasoline prices was to conserve energy and to pressure major oil-exporting countries such as Saudi Arabia to produce more oil for the world market. He stoutly opposed relaxing environmental standards to make it easier to drill for oil in the United States, and he insisted that the Arctic National Wildlife Refuge be kept closed to oil exploration.

The partisan division among the two parties' presidential candidates is mirrored in public opinion

data. Figure 7 shows the partisan breakdown of support for the proposal to deal with the energy situation by “oil exploration in the Alaskan Arctic Wildlife Refuge.” Republicans supported the idea by a 54 to 43% margin, but Democrats opposed it by a 65 to 27% margin. People who did not identify with either political party—that is, those who call themselves independents—fall between Republicans and Democrats in their policy views.

Similar partisan differences exist across many energy-related and environmental issues. The same May 2001 Gallup poll found sharp differences regarding whether to invest more money in natural gas pipelines (favored by 76% of Republicans but only 54% of Democrats), whether to drill for natural gas on federal land (favored by 75% of Republicans but only 55% of Democrats), and whether to increase the use of nuclear power (favored by 59% of Republicans but only 45% of Democrats). Other national and state polls routinely find similar partisan differences, with Republicans typically comprising the group that prefers conventional energy development and Democrats comprising the group that prefers conservation and alternative energy sources.

Another political orientation that helps explain public opinion on energy and environmental issues is ideological self-identification—that is, whether people consider themselves to be liberal, moderate, or conservative. Ideological self-identification is a curious characteristic. Pollsters ask respondents a question such as “Generally speaking, in politics do you consider yourself as conservative, liberal, middle-of-the-road, or don’t you think of yourself in those terms?” The terms liberal and conservative are

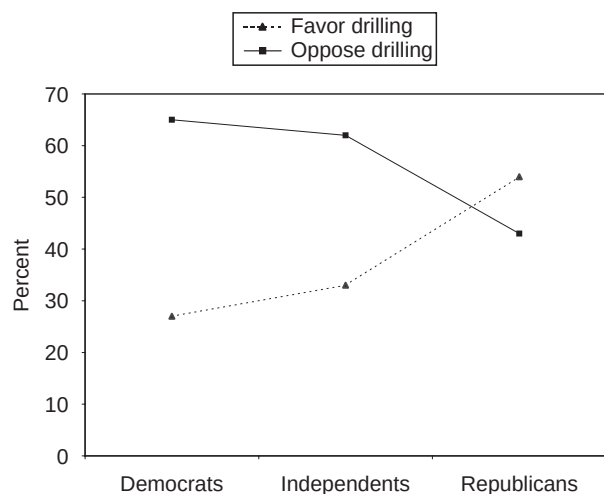


FIGURE 7 Support for oil drilling in the Arctic National Wildlife Refuge by party. Source: Gallup poll, May 2001.

in wide use in public debate. However, repeated studies for more than 50 years show that only approximately half the public can offer even crude definitions of what the labels mean. Typically, approximately one-third of respondents say “I don’t know” when asked for a definition. Another one-sixth will offer definitions that are not even remotely recognizable. Nevertheless, the ideological labels that people give themselves predict their opinions relatively well, including their opinions on many energy issues.

Environmental problems emerged as political issues relatively recently in comparison with traditional economic conflicts over labor laws, welfare systems, and tax codes. Nevertheless, when they did begin to draw the public’s attention in the 1960s, liberal leaders argued for laws imposing clean air standards, clean water standards, and other pro-environmental regulations. Conservative leaders argued against allowing the government to impose new regulations on businesses. Environmentalism quickly became an established part of liberalism, whereas pro-development views became part of the conservative agenda. This liberal-conservative distinction on environmental issues spread quickly among both political leaders and the public.

Figure 8 shows the ideological divisions over the dispute about offshore oil and gas drilling along the California coast. Seventy-seven percent of the people who say they are “strong liberals” oppose more drilling, whereas only 34% of strong conservatives oppose it. On the pro-development side, the figures

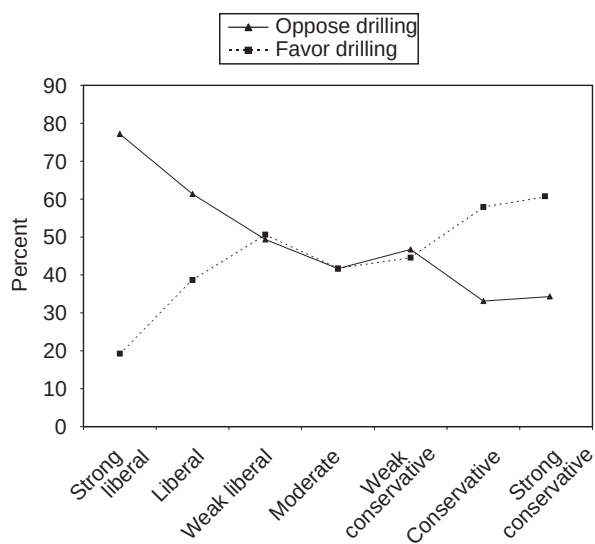


FIGURE 8 Californians support for offshore oil drilling by ideology. Source: Field Institute poll, May 2001.

are reversed, with 61% of strong conservatives but only 19% of strong liberals favoring more drilling.

Taken together, party identification and ideological self-identification do a much better job of explaining public opinion on energy issues than do demographic variables. Nevertheless, opinions on environmental and energy-related issues are not understood or explained as well by public opinion scholars as opinions in other areas (e.g., welfare and civil rights). Given the importance of public opinion to policy makers, research in this area continues.

SEE ALSO THE FOLLOWING ARTICLES

Energy Development on Public Land in the United States • Hydropower Resettlement Projects, Socio-economic Impacts of • Lifestyles and Energy • Media Portrayals of Energy • Public Reaction to Electricity Transmission Lines • Public Reaction to Nuclear

Power Siting and Disposal • Public Reaction to Offshore Oil • Public Reaction to Renewable Energy Sources and Systems

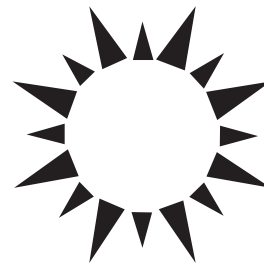
Further Reading

- Dunlap, R. E., and Mertig, A. G. (eds.). (1992). "American Environmentalism: The U.S. Environmental Movement, 1970–1990." Taylor & Francis, Philadelphia.
- Freudenburg, W. R., and Rosa, E. A. (eds.). (1984). "Public Reactions to Nuclear Power: Are There Critical Masses?" Westview, Boulder, CO.
- "The Gallup Poll Monthly." Gallup Organization, Princeton, NJ.
- Jones, R. E., and Dunlap, R. E. (1992). The social bases of environmental concern: Have they changed over time. *Rural Sociol.* 57, 28–47.
- Kempton, W., Boster, J. S., and Hartley, J. A. (1995). "Environmental Values and American Culture." MIT Press, Cambridge, MA.
- Richman, A. (1979). The polls: Public attitudes towards the energy crisis. *Public Opin. Q.* 43, 576–585.
- Smith, E. R. A. N. (2002). "Energy, the Environment, and Public Opinion." Rowman & Littlefield, Boulder, CO.

Public Reaction to Nuclear Power Siting and Disposal

EUGENE A. ROSA and JAMES RICE

Washington State University
Pullman, Washington, United States



-
1. Background
 2. Nuclear Generation
 3. Evolving Contexts
 4. Public Reactions to Nuclear Power in the United States
 5. Environmental Consciousness and Nuclear's Fortunes
 6. The Present and the Future
 7. Nuclear Imagery
 8. Nuclear Waste
 9. Conclusion

Glossary

energy The capacity for doing work. Energy may be stored, and it may be transferred from one form into another. Forms of energy include electrical, chemical, thermal, and mechanical.

global warming Atmospheric gases such as carbon dioxide and other greenhouse gases make the earth warmer than would direct sunlight alone by trapping heat from the sun as it is reradiated off the earth's surface and absorbed in the atmosphere rather than escaping into space.

greenhouse gases Atmospheric gases that trap the heat of the sun in the earth's atmosphere. Carbon dioxide, methane, nitrogen oxides, and ozone are common examples.

nuclear energy Power obtained by splitting (fission) or joining (fusion) atoms of radioactive materials, such as uranium.

nuclear waste The materials, both liquid and solid, left over from the production of nuclear energy. Nuclear waste exhibits varying amounts of radiation and can be harmful to humans.

power The rate at which energy flows.

radiation The flow of energy via electromagnetic waves. Radiation is produced by radioactivity.

radioactivity Produced when nuclei are undergoing transformation and emitting radiation.

reactor A device in which a controlled nuclear chain reaction can be maintained to produce heat. The heat is then typically used to produce steam, which in turn is

used to run conventional turbine generators to produce electricity.

uranium A radioactive element found in ore. Uranium atoms can be split to create energy.

The harnessing of the atom, opening the way for the generation of electricity with nuclear energy, was met with a technological enthusiasm perhaps unmatched in modern history. Nuclear energy was a technology with unprecedented potential. Early in its development, nuclear power was claimed to be the cheapest form of electricity and was projected to provide a substantial portion of the electricity needs in the advanced nations of the world. History, however, conspired to stall the march of nuclear power with the result that these projections were never realized despite the technology's continuing potential. Standing in the way of its destined path to triumph, a variety of unanticipated technological, social, and political obstacles cloud the technology's future nearly everywhere. This article examines the key social obstacles leading to the stalemating of nuclear power.

1. BACKGROUND

Although born of war, nuclear fission technology was soon converted to the "peaceful" generation of electricity. The emergence of nuclear power as an energy source and the high hopes it engendered were hardly surprising in an evolutionary context. Passing through stages from early to mature industrialization was synonymous with a substitution process—the substitution of more readily abundant, less expensive, energy potent fuels for dwindling, more expensive, less potent fuels. With the appearance of fission technology, the generation of commercially produced electricity seemed the rightful heir to propel this evolutionary process forward. After all,

the technology could convert puny amounts of raw fuel into usable energy so vast that it would be “too cheap to meter,” displacing the reliance on fossil fuel electricity in the process. However, standing in the way of nuclear power’s destined path to success were a wide variety of unanticipated technological as well as social and political obstacles that cloud the technology’s future nearly everywhere.

The tension between the technology’s potential and its public acceptance was captured in the most repeated quote in nuclear history—by Alvin Weinberg, the doyen of nuclear power:

We nuclear people have made a Faustian bargain with society. On the one hand, we offer—in the catalytic burner—an inexhaustible source of energy. But the price that we demand of society for this magical energy source is both vigilance and a longevity of our social institutions that we are quite unaccustomed to.

Another price demanded of any centralized technology promoted by the national state is acceptance by its citizens. Public reactions to nuclear technology in the United States and elsewhere produced a checkered history for the technology that evolved in five overlapping phases: an era of initial enthusiasm, a rejection at the local level, a rise in general public concern and the assertion of alternative technologies, a period of ambivalence, and the current era of persistent pragmatism. Earlier phases prefigured later ones. Therefore, the evolution of public opinion about nuclear power in the United States can be understood as the dynamic and historically contingent intersection between the actions of the federal government, private utilities, industrial contractors, social movement organizations, and the general public in relation to this complex technological system perceived as promising but risky.

The decade of the 1960s, the earliest phase of nuclear power development in the United States, was unquestionably one of unbridled optimism, based, as it was, on the promise of nearly unlimited supplies of cheap electricity. During this period the federal government crystallized its roles as both promoter (with massive investments in R&D and industry subsidies that may total more than \$1 trillion) and as regulator of the technology—roles it continues to fulfill today. The eventual recognition of the inherent conflict of interest in the two roles led to their separation in 1974 by the creation of the Energy Research and Development Administration to fulfill the promotion role (in 1977 to be absorbed into the newly created Department of Energy) and the Nuclear Regulatory Commission to fulfill the reg-

ulatory role. During this optimistic phase, the nuclear industry was most vigorous in its promotion of the technology. For example, beginning in 1963, plants were offered to utilities for the first time on a strictly commercial basis with so-called “turnkey” contracts, so named because the purchasing utility was guaranteed that the plant would be completed at a fixed price and needed to do little more than “turn the key” to begin operations.

As the initial developer of nuclear technology, the United States became the world leader in efforts to build nuclear power plants, to commercialize them, and to promote the technology among the industrial nations. It continues to exercise major influence on the industry worldwide. Also, its commitment to nuclear, with only slight variation, has persisted through 10 successive presidential administrations. Hence, this early and lengthy experience with nuclear energy makes the United States an especially suitable focal point for examining the evolution of public acceptance of the technology. Although this examination focuses principally on the U.S. experience, it closely reflects the experience of other countries, such as Germany, Canada, Japan, the United Kingdom, and leading promoter France, although France exhibits certain deviations from overall patterns.

2. NUCLEAR GENERATION

In January 2003, 31 countries worldwide produced electricity with nuclear energy in 441 power reactors. World generating capacity totaled 359 GW (359,000 MW), representing 16% of total worldwide electricity production. With the completion of the 32 plants currently under construction, generating capacity is expected to increase by 7.5%, and it may increase another 10% if the 32 additional planned plants (35% larger, on average) are eventually completed.

Table I presents data for each of the world’s nuclear nations. In terms of total generating capacity, the United States leads the world with 98.2 GW (104 reactors), followed by France with 63.1 GW (59 reactors), Japan with 44.3 GW (54 reactors), Germany with 21.3 GW (19 reactors), Russia with 20.1 GW (30 reactors), South Korea with 14.9 GW (18 reactors), the United Kingdom with 12.1 GW (31 reactors), Ukraine with 11.2 GW (13 reactors), and Canada with 10 GW (14 reactors). As a percentage of total electricity, Lithuania leads the way with more than 80% of its electricity produced by nuclear, followed closely by France (the largest

TABLE I
Nuclear Capacity and Operations of the World's Nuclear Nations

Country	Nuclear generating capacity (Gwe)	No. of nuclear power reactors (as of February 2003)	Nuclear as a % of total electricity generation (as of 2002)
Argentina	0.935	2	7.2
Armenia	0.376	1	40.5
Belgium	5.760	7	57.3
Brazil	1.901	2	4
Bulgaria	2.722	4	47.3
Canada	10.018	14	12.3
China	5.318	7	1.4
Czech Republic	3.468	6	24.5
Finland	2.656	4	29.8
France	63.073	59	78
Germany	21.283	19	29.9
Hungary	1.755	4	36.1
India	2.503	14	3.7
Japan	44.287	54	34.5
South Korea	14.890	18	38.6
Lithuania	2.370	2	80.1
Mexico	1.360	2	4.1
The Netherlands	0.450	1	4
Pakistan	0.425	2	2.5
Romania	0.655	1	10.3
Russia	20.793	30	16
Slovak Republic	2.408	6	54.7
Slovenia	0.676	1	40.7
South Africa	1.800	2	5.9
Spain	7.574	9	25.8
Sweden	9.432	11	45.7
Switzerland	3.200	5	39.5
Taiwan	4.884	6	21
Ukraine	11.207	13	45.7
United Kingdom	12.052	31	22.4
United States	98.230	104	20.3
Total	358.46	441	30 ^a

^a 16% of world total.

exporter of nuclear energy) at approximately 80%, Belgium at 57%, Ukraine at 46%, Sweden at 45%, Bulgaria at 47%, Switzerland at 40%, Japan at 34%, South Korea at 38%, the United States at 20%, Canada at 12%, and Russia at 16%.

Generating 16% of all electricity consumed in the world, nuclear power clearly remains a significant

energy source. However, the significance diminishes considerably when placed in the context of projections made in the 1960s and 1970s during the heyday of the technology. One estimate projected 3000 plants worldwide by 2000 (nearly 7 times actual capacity), whereas another estimate, the 1972 Atomic Energy Commission's most optimistic estimate for the United States alone, called for 1200 plants by the same year (nearly 12 times actual capacity). Not a single plant was ordered in the United States after 1978, and every plant ordered after 1974 was eventually canceled so that not a single reactor built in the United States was ordered after 1973.

A number of industrial nations have chosen to forego the nuclear option, especially in the post-Chernobyl era: Australia, Austria, Denmark, Italy, Ireland, and Norway. Sweden continues to operate its 11 nuclear reactors (1 was closed in 1999) and has extended the 2010 date calling for the shutdown of all plants due to a 1980 national referendum, but it continues its policy of phasing out nuclear power. Germany, an early enthusiast for nuclear energy, passed legislation in 1998 for the eventual phase out of nuclear power but subsequently modified the policy in 2001 to limit the operational lives of existing plants to 32 years, thereby deferring any immediate closures. Japan, 80% dependent on imported energy, has scaled back its design for nuclear because of a September 1999 accident in Tokai, Mura, Ibaraki Prefecture, where two workers died from severe radiation and dozens of others were exposed.

A clear exception to this general pattern for the trilateral nations is Finland, the first nation in the world to have an approved site for the disposal of high-level nuclear waste and whose parliament in mid-2002 approved the building of a fifth nuclear power plant. Finland's success with its nuclear program may be the harbinger for a jump-start of the sagging nuclear programs of the majority of Western nations, or it may be an anomaly stemming from other unique features of Finnish history (e.g., its recent transformation from an agrarian to industrial society) and its unitary form of democracy.

Apart from Japan, nuclear capacity continues to grow in Asia (at the world's fastest rates), especially in China, the Republic of Korea, Taiwan, India, and Pakistan. As of 2002, more than two-thirds of the more than 32 reactors in 11 countries under construction throughout the world were located in Asia. Whether these nations will continue to experience growth uninterrupted by social and political obstacles or whether they will encounter difficulties similar to those experienced by Western

nations will be the subject of considerable scrutiny throughout the world.

In summary, although nuclear energy has assumed a significant and durable role in the energy mixes of a number of nations, its share of that mix is far short of the projections of three decades ago and with current technology it is projected to grow only moderately or even decline by midcentury. Furthermore, nuclear power is stalemated in many countries. What accounts for the striking underperformance of this technology? A number of factors conspired to derail nuclear's trajectory to predicted success: shifting energy markets, geopolitics, capital shortages, historical turns, and a failure to gain widespread public support. To what extent will these factors persist to obstruct a nuclear future?

A proper assessment of the future prospects for nuclear power requires consideration of those features of the historical context that are undergoing change. Two features loom particularly large in this respect, an internal one and an external one. Recognizing fundamental flaws in past reactor designs, such as the mushrooming of alternative designs and plant sizes and of limitations in safety systems, the world nuclear industry is working toward the development of generation IV reactors, a series of designs that are simpler, more fuel efficient, and arguably safer. The second, larger context is the scientific consensus over the warming of the planet due to the accumulation of greenhouse gases, especially CO₂. The evolution of these contexts is discussed in the next section.

Despite past experience and what appears to be a propitious new environment for nuclear power, the future of the technology will rest, ultimately, on public acceptance. Hence, the remainder of this article is devoted to the evolution of public attitudes toward nuclear power and to the implications the historical patterns have for the future of the technology. It is organized around three general themes: the evolutionary status of public attitudes toward the technology, the willingness by publics to see nuclear growth, and how publics view the technology abstractly.

3. EVOLVING CONTEXTS

3.1 New Reactor Designs

Nuclear power reactors are typically categorized in generational terms, reflecting evolutionary changes in design, size, and operational uses. Generation I reactors, the earliest to be developed in the 1950s and 1960s, were typically small (often based on

submarine designs), with few still in operation today. Generation II reactors are the most common type of reactor in use today, dominated by U.S. light-water technology. Generation III reactors, an advanced design, are still in their incipient stages of development, with only a handful in operation in Japan. Generation IV reactors are the most advanced designs currently under development by 12 nuclear nations.

Comprising more than six different designs and systems, generation IV reactors all have in common higher (often much higher) operational temperatures than those of reactors currently operating. For some designs, this will allow for hydrogen production. Among the many other benefits arguably resulting from this generation of reactors are simpler designs and operations, thereby increasing safety; improved fuel efficiency, thereby improving the competitiveness of nuclear electricity; and use of a closed fuel cycle, thereby reducing the amount of high-level wastes to be managed.

Whether the increased safety of this new generation of reactors will attract widespread support for nuclear power will depend on a safety record that confirms the safety claims and whether publics trust the proponents of the new designs.

3.2 Global Climate Change

The emergence of a scientific consensus over global climate change has shed a new spotlight on nuclear power. Because it does not emit the principal greenhouse gas, CO₂, nuclear power is now viewed by many observers as a principal source of future energy poised for renewed, rapid growth. If there were dramatic new growth in nuclear power it could indeed reduce the rate of CO₂ emissions, but it should be recognized that nuclear is not the sole solution to global warming. Nuclear energy currently accounts for only 2.5% of total world energy consumption (8% in the United States and 33% in France), and its growth is limited by electricity's meager 16% share (the only type of usable energy produced by nuclear) of world energy consumption. The 16% share could increase as a function of electricity's growing fraction of total energy use, but nuclear's potential for addressing climate change issues nevertheless remains relatively small.

Whether public recognition of the seriousness of climate change and nuclear's ameliorative role in curbing it will swing the pendulum of public opinion to majority support is a difficult question to assess. Among the many unresolved issues the question raises is how citizens will balance a very diffuse and

perhaps distant risk whose observable consequences are not immediately known (climate change) against a very proximate risk (nuclear accidents) for which some of the consequences are known. The social psychological literature suggests that the proximate risk would trump the diffuse one.

4. PUBLIC REACTIONS TO NUCLEAR POWER IN THE UNITED STATES

4.1 Long-Term General Trends

4.1.1 Building More Nuclear Power Plants

As noted previously, the ultimate fate of nuclear power rests with public acceptance. The longest running nuclear time series, providing the best available window for assessing the evolution of public opinion, concentrates on attitudes toward the continued growth of nuclear power. Similar, but not identical, growth questions have been asked by leading polling organizations Louis Harris and Associates and Cambridge Reports Research Inter-

national. The evolution in opinion about nuclear plant construction, including opinion reversal, can best be seen by charting trend lines for these data series. Figure 1 presents a trend line from the Harris data and Fig. 2 presents a trend line from the Cambridge data.

The Harris item asking for opinions about building more nuclear plants was first asked in 1975 and repeated until 1990. In general, until the late 1970s a majority of Americans supported, often by wide margins, growth in nuclear power (an exception is the April 1976 result showing a plurality of support). Then, immediately following the accident at Unit 2 at Three Mile Island (TMI) in Pennsylvania in April 1979, this support eroded considerably but still consistently exceeded opposition. With the May 1979 survey, support had regained its majority position but appeared to be standing on much shakier ground than before the accident: Over the course of 12 surveys conducted over the next 22 months, support for nuclear power, although generally consistent, vacillated between bare majority and plurality. In one case, May 1980, opposition actually exceeded support by

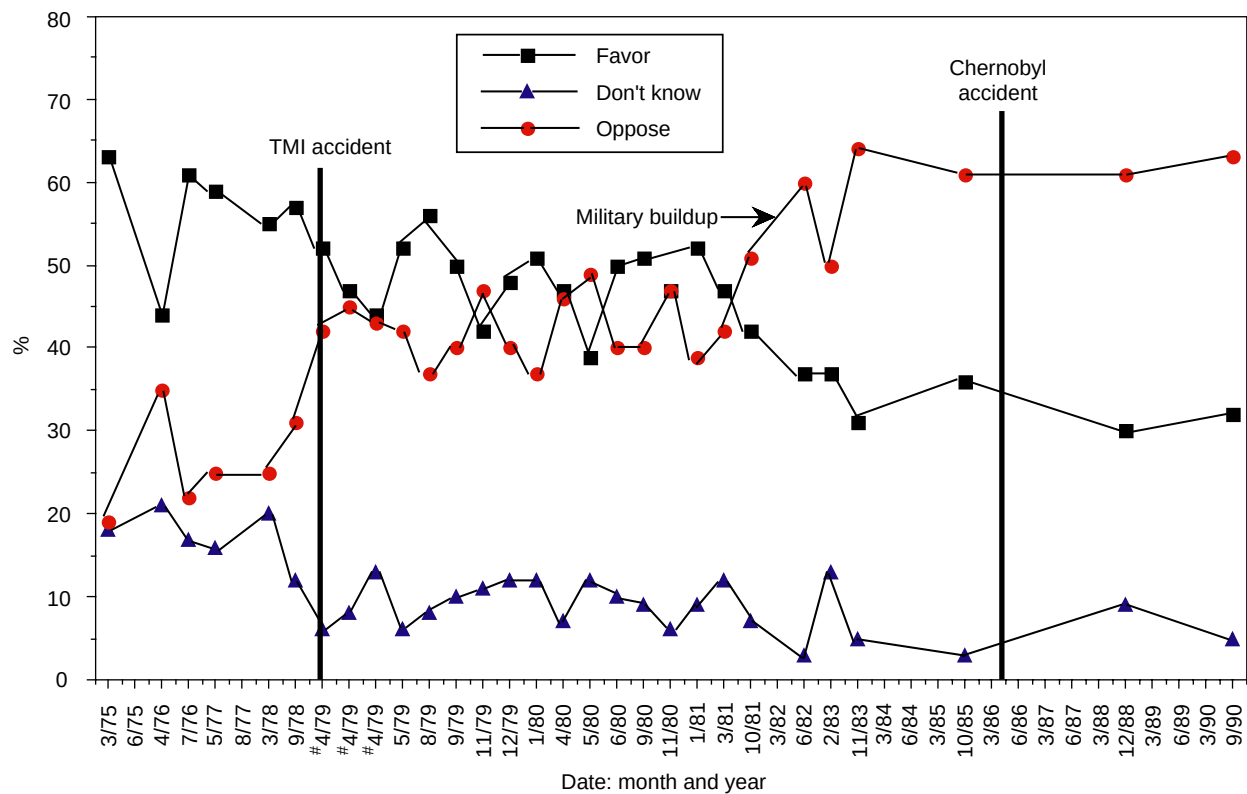


FIGURE 1 Public attitudes toward building nuclear power plants in the United States. Question: "In general do you favor or oppose the building of more nuclear power plants in the United States?" TMI, Three Mile Island. #Indicates multiple surveys in the same month. Source: Louis Harris and Associates national samples.

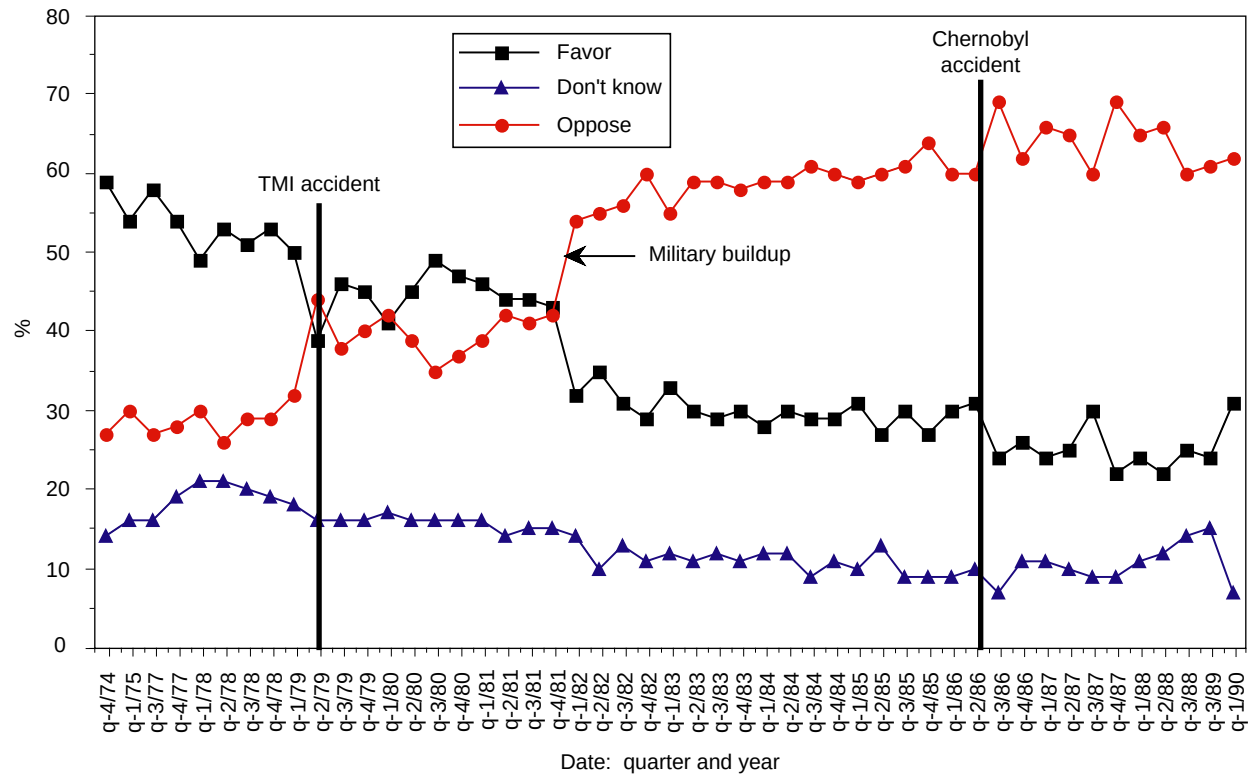


FIGURE 2 Public attitudes toward building nuclear power plants in the United States. Statement: “I am going to read you a list of proposals for dealing with the energy crisis, and I’d like you to tell me whether you generally favor or oppose each one ... Building more nuclear power plants.” TMI, Three Mile Island. Source: Cambridge national samples.

10 percentage points, but this was a temporary setback lasting for only a short period of time.

In contrast, the October 1981 result showing majority opposition signaled not just a temporary setback but also a break with the past—indeed, a reversal in public preference for nuclear growth. Since then, opposition has consistently exceeded support. In fact, over the past 4 surveys spanning 7 years, opposition exceeded support by a margin of approximately 2:1, a complete reversal from the earliest Harris survey.

Cambridge, likewise, queried the public repeatedly about whether to build more nuclear power plants. In the fourth quarter of 1974, Cambridge began asking Americans their views about building more nuclear power plants as a way of dealing with the energy crisis. That Cambridge contextualized its question within a concern for the energy crisis is perfectly understandable given that it first polled the public on the subject in the months following the Arab oil embargo of 1973–1974 when national agendas were dominated by the idea of an “energy crisis.” This contextualization distinguishes the Cambridge series from the Harris one, and one might expect it to have

a noticeable affect on the overall trends. With the notion of crisis in mind, citizens might be willing to accept any energy source that would mitigate or eliminate the crisis, including nuclear power. It might also be the case that in later surveys, long after the idea of crisis had receded from public consciousness, the effect could be in the opposite direction: Because a crisis no longer exists, there is no urgent need to build more nuclear power plants.

Figure 2, presenting the Cambridge trends, exhibits similar patterns to the Harris results; nuclear power enjoyed consistently strong support throughout the 1970s, at least up until the TMI accident in April 1979. The effect of the accident on public attitudes is apparent, as opposition exceeded support for the first time in the second-quarter 1979 survey. By the third quarter of that year, however, support had rebounded sufficiently to outpace opposition, but not by as large a margin as before the accident. Furthermore, while support remained consistent during the next $2\frac{1}{2}$ years, covering 10 surveys, it was only a plurality of support, and sometimes barely that. Then, sustaining the pattern of the Harris results, a sharp reversal in opinion appeared

(showing up for Cambridge in the first quarter of 1982), with a majority of respondents expressing opposition to further nuclear plant construction. What is most noteworthy about this shift from majority support to majority opposition is the fact that it endured across all subsequent surveys over the course of 8 years. Furthermore, for most surveys the ratio of opposition to support approaches or exceeds 2:1—again, a near-perfect reversal of public preference in the early 1970s. Finally, the substantive similarity between the Harris and Cambridge results suggest that wording effects, if there are any, do not markedly alter the overall patterns.

In general, despite the slight wording changes between the polling organizations, a common and persistent pattern of growing opposition to the building of more nuclear power plants emerges. The pattern comprises three distinct stages: an early stage in the 1970s when Americans were enthusiastic about nuclear growth, a second stage of ambivalence following TMI when a less enthusiastic plurality of citizens consistently supported growth, and a third stage emerging in the early 1980s when a decisive majority of Americans opposed building more nuclear power plants in the country. Apace with the decline in support for nuclear power plants was the decline in the proportion of “don’t knows,” reflecting, perhaps, the eventual crystallization of attitudes.

A focus on the timing of the reversal illuminates still another striking feature in the overall pattern: the change in attitudes, with opposition overtaking support, occurred more than 4 years before the Chernobyl nuclear accident in the Ukraine. This was apparently due to a renewed fear of nuclear war stimulated by the Reagan administration’s bellicose posture toward the Soviet Union and the accompanying buildup of nuclear arms. Whatever the reason for the striking shift in attitude, it set in motion a steady growth in opposition that continued to the very end of each data series. With such a high level of opposition, there was little room for further opposition, so the Chernobyl accident evidently did not have a marked impact on public attitudes: At the time of the accident, opposition already exceeded support (best seen in the Cambridge results in Fig. 2) by nearly two to one. Whatever effects Chernobyl may have had, they were apparently negligible compared to the sustained long-term decline in support for new nuclear power plants.

Although these data provide a clear picture of the evolution of opposition to nuclear power, the question about building more nuclear plants has not been repeated for more than a decade. Given the propi-

tious new winds for nuclear power noted previously, how can we be certain that opposition to the building of more plants remains at this high level? Four polls address this question. A March 1999 poll of registered voters found 60% opposed and 26% in favor of building more nuclear power plants. A March 2001 Gallup poll of all adults found less but still majority opposition (51 vs 44%). Two 2001 polls of all adults by ABC News/*Washington Post* sustain the conclusion that opposition to the building of nuclear power plants is still in the majority. In its January poll, 60% were either somewhat or strongly opposed, whereas 37% were somewhat favorable or strongly favorable. Its June poll revealed some relaxing of opposition but still with a majority (52%) opposed, whereas 41% were in favor.

Clearly, the American public does not support the construction of additional nuclear power plants. This is the unmistakable message contained in the longest running time series assessing nuclear opinion, reaffirmed by recent survey data.

The public opinion experience of other countries deviates somewhat from the pattern in the United States but available data suggest that citizens in most nuclear nations have lingering reservations about the technology. For example, although opposition to nuclear power in Sweden has softened during the past two decades, opposition still outpaces support for the technology. Available public opinion data in Japan show a very similar pattern to that of the United States, except that opposition’s overtake of support did not occur until the Chernobyl accident in 1986, and recent data show the Japanese to be more fearful of a nuclear accident than any other nation.

In comparison to the nuclear stalemate in much of the world, France has experienced nearly uninterrupted success with its nuclear program. This has repeatedly prompted nuclear proponents to point to France as the exemplar of not only a successful nuclear program but also a model for other nuclear nations to emulate. Whether France is an appropriate model for other nations to follow is subject to considerable debate. Key differences distancing France from other nuclear programs include its monopoly ownership by a single government utility, Electricité de France; its unique system of training for its technological elite; its strategy of adopting a single design for all its reactors; its practice of siting multiple reactors at a single location; its unitary democratic structure; and its policy style (called technocratic etatism), in which individual liberties are subordinated to the will of the state.

Added to these obvious differences are others more subtle but no less important to France's nuclear success. In public opinion polls, French citizens are consistently more supportive of nuclear than their American counterparts. What accounts for this persistent pattern? In the only comparative opinion study between French and American publics it was found that perceptions of the risk of nuclear power were virtually identical between the two publics, thereby eliminating risk as an explanation for the national differences. French citizens distance themselves from Americans in their strong belief in the need for nuclear power, their more fatalistic viewpoint about its risks, their greater trust of nuclear management, and especially their greater trust in experts and civil authority (a feature of etatism). There are likely similar differences between the French and citizens of other nuclear democracies.

4.2 Proximate Trends

4.2.1 *Nuclear Power as a Source of Electricity*

The sole commercial use of nuclear power is the generation of electricity. Despite the long-term trends showing a growing dissatisfaction with the building of more nuclear power plants in the United States, perhaps the public is more supportive of the technology when it is connected directly to its instrumental purpose: the generation of electricity. In a poll taken soon after the Chernobyl accident, the public was evenly split on the use of nuclear power to generate electricity, but a strong 55% majority favored nuclear in a follow-up poll in 1989. Other polls on public attitudes toward nuclear power as an electricity source followed. For all three polls conducted between 1992 and 1993, a majority of respondents favored (strongly or somewhat) the use of nuclear power to generate electricity. It should be noted, however, that these polls did not ask for unqualified support for nuclear power but as one of the ways of providing electricity. Americans are apparently aware that nuclear is already an important energy source for the generation of electricity. Two surveys in 1986 and 1989 asked how important nuclear was in meeting the nation's electricity needs today. A sizable majority in both (70%) agreed that it was important. The most recent data show a considerable softening of support, although the general pattern is sustained.

4.2.2 *Local Siting of Power Plants*

It is one thing to have an opinion toward the construction of nuclear power plants in the abstract;

it is another thing to be confronted with the prospect of having a plant built nearby. Figure 3 presents trend data from national samples that asked about having the local electric company build a nuclear power plant in the area.

The early siting of nuclear power plants was hardly a controversial matter. Indeed, in five of the earliest polls 55% responded favorably, nearly twice the level of opposition. However, opposition to local nuclear facilities began to increase as early as 1974. By 1976, a slight plurality, and by 1978—before the TMI accident—a clear majority of respondents opposed the construction of a nuclear power plant within 5 miles; in less than a decade the opposition reached 73%. Polling asking whether citizens would be willing to have a nuclear power plant within 10 miles of their community attracted 54% opposition in 1999 (compared to 21% in favor) and 50% opposition in 2000 (compared to 22% in favor). In a 2001 poll, 63% of the adults sampled were somewhat or strongly opposed to the construction of a nuclear power plant “in their area,” whereas 33% were somewhat or very favorable to the idea.

It might be expected that, as had been the case regarding the building of nuclear power plants in the abstract, Americans would be more tolerant of local nuclear plants if they were reminded that the purpose of the technology was to provide electricity. Cambridge and Gallup, using an identically worded item, attempted to assess this possibility in a series of polls from 1986 to 1990. The most popular of three possible responses, attracting near-majority to clear-majority support, was to “reserve judgment”—perhaps reflecting, on the one hand, a reluctance to reach a firm position until apprised of the specific circumstances surrounding such a decision and, on the other hand, a willingness to leave open the nuclear option as a hedge against future electricity shortages. Of those who had evidently formed an opinion about a local plant, opposition consistently outweighed support, but by a declining margin over time.

4.2.3 *Local Host Communities*

As suggested by the data (Fig. 3) about the local siting of nuclear power plants, in the early years of nuclear commercialization communities hosting plants were quite supportive of the technology. Indeed, a number of communities actively sought to attract nuclear power plants for economic reasons and as a matter of technological pride. This early enthusiasm is sustained by the trend data in Fig. 4, a summary of 36 host community surveys.

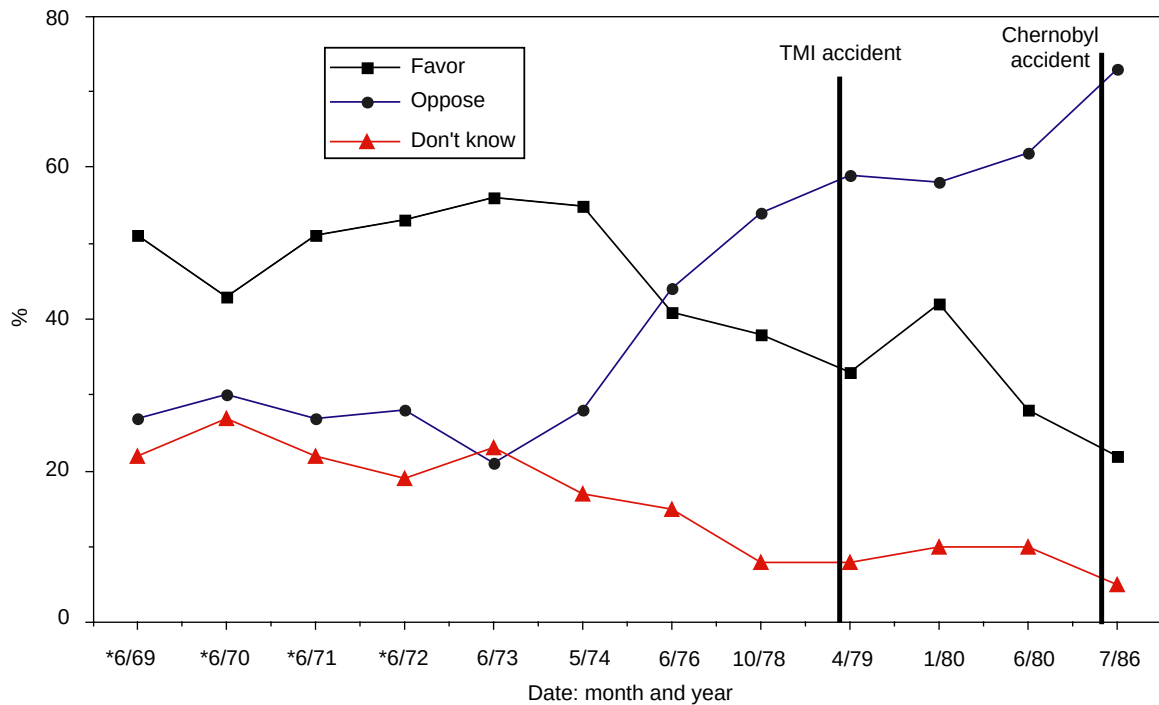


FIGURE 3 Public attitudes toward construction of local nuclear power plants. Question: "Suppose your local electric company said it wanted to build a nuclear power plant in this area. Would building such a plant be all right with you, or would you be against it?" TMI, Three Mile Island. *, Month survey conducted not reported in original sources. Midyear dates were assumed. Source: Cambridge national samples. Boulder, CO.

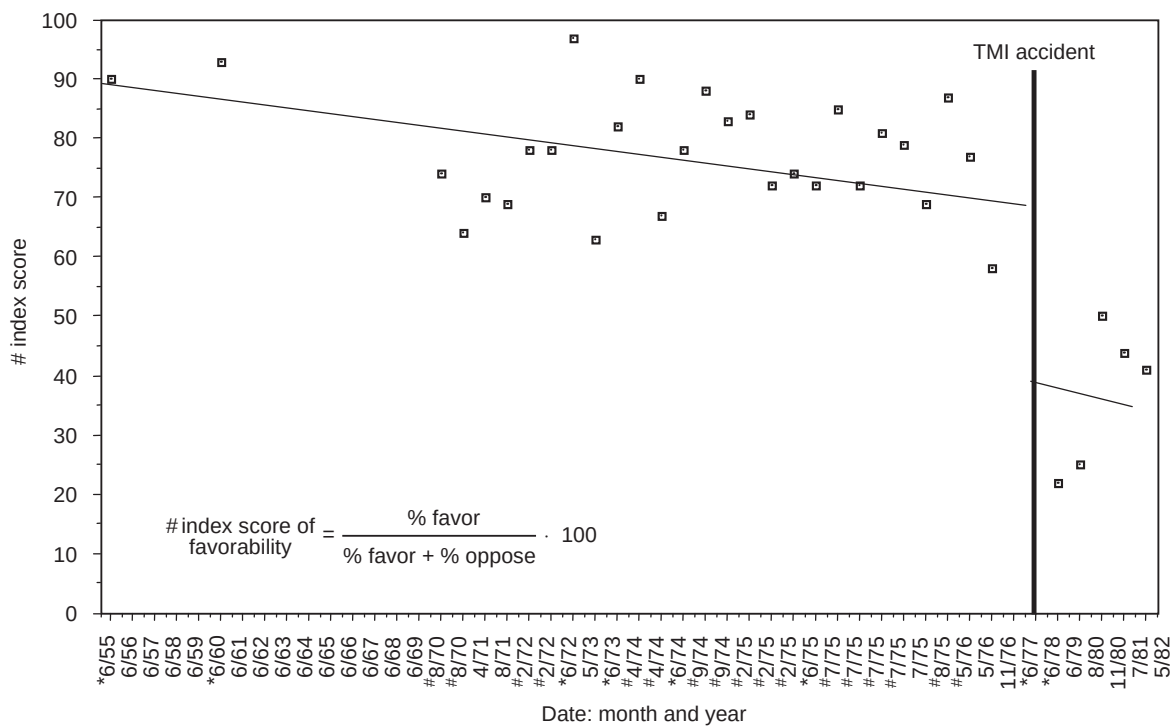


FIGURE 4 Host community attitudes toward nuclear power plants before and after the Three Mile Island (TMI) accident. *Month survey conducted not reported in original sources. Midyear dates were assumed. #Indicates multiple surveys in the same month. Source: Cambridge national samples.

Post-TMI data suggest, however, that host community support can no longer be taken for granted. In none of the surveys conducted before the TMI accident was opposition found among more than 33% of respondents; in no known survey after the accident has there been less than 50% opposition. Regression analysis shows that more than 70% of the explained variance across the surveys is due to the TMI accident. Host communities in the United States, once supportive of nuclear facilities in their vicinity, now show considerable reservation about having those facilities nearby. Also, the siting of nuclear facilities is a persistent problem in a number of nuclear nations.

5. ENVIRONMENTAL CONSCIOUSNESS AND NUCLEAR'S FORTUNES

The presented trend data clearly illustrate the dramatic impacts key events (the TMI accident and military buildup) had on public opinion about nuclear power. However, preceding the reversal in general attitudes by several years was growing disenchantment with local nuclear facilities—the principal fault line in the growth of nuclear power. What accounted for this early disenchantment?

In the late 1960s and early 1970s, the Western world experienced a profound cultural change in collective consciousness in which environmental impacts, previously a set of disparate concerns, coalesced into an environmental consciousness. Signature features of this change were the passage of the National Environmental Protection Act in 1969 in the United States, a model for similar legislation elsewhere, and the first Earth day in 1970, the largest public demonstration in the history of the planet—attracting many millions of participants worldwide. Although these events did not immediately lead to a reversal in nuclear's fortunes, they established a cultural framework of unprecedented scrutiny of the technology. Early local opposition to nuclear power plants reflected the emergent consciousness with environmental issues such as thermal pollution (release of heat to rivers and lakes) and threats to marine species as the rallying cries for protest.

When the accidents at TMI and Chernobyl took place, publics everywhere were predisposed to see the events in the most negative light. Another element of the new consciousness, later to impact nuclear directly, was ever-greater demands on policy makers

to more actively engage the public in decisions regarding large-scale technologies.

The emergence of a consciousness potentially critical of nuclear energy was one side of the energy supply coin. Missing was the other side of the coin that provided a coherent alternative, a revolutionary vision of an energy supply mix absent a nuclear component. In the 1970s, many scholars were attracted to the issue, but none could match the impact of E. F. Schumacher's "appropriate technology" vision or Amory Lovins's "soft path" vision. Now, instead of a focus on nuclear technology versus its alternatives, the focus was broadened to a criticism of the megascale, centralized energy supply systems of industrial societies. Lovins went further, developing a detailed blueprint of renewable energy systems absent a nuclear component.

6. THE PRESENT AND THE FUTURE

6.1 Pragmatism

Does the sizable opposition to the building of more nuclear power plants translate into a deep-seated aversion to the technology, into an unreflective rejection of all matters nuclear or even a willingness to get rid of the technology completely? Is the opposition the result of sensible caution about a large-scale technology that is neither simple nor risk-free? In one survey, respondents were asked whether nuclear power was a good choice, a realistic choice, or a bad choice. Strong pluralities to majorities consistently said it was a realistic choice. The remaining 40–50% of the public was more likely to say it was a bad choice, but sometimes by only narrow margins.

With regard to eliminating nuclear power, only minorities of Americans have supported this option. The most frequently chosen option, attracting sizable pluralities to majorities, favors the status quo: Let the existing nuclear plants operate, but do not build any more. Taken together, the picture that emerges from these items is not one of a knee-jerk public passionately committed to the riddance of nuclear energy but of a pragmatic public weighing technological risks against unforeseeable future energy needs. As a hedge against a shortfall in future energy supply, the public appears willing to leave the nuclear option open.

6.2 Nuclear Futures

What does the public believe the future holds for nuclear power? The longest running time series, the

Harris and Cambridge ones, asking about building more nuclear plants are of limited use in addressing this question because the time orientation of these items is ambiguous. The ambiguity can be avoided with questions that specifically refer to the future, as have a number of polls.

Since 1985, Americans have repeatedly been asked how important they think nuclear energy plants will be in meeting the nation's electricity needs in the years ahead. The trend data of responses to this question are presented in Fig. 5.

It is clear that sizable majorities of Americans see an important role for nuclear power in the future. Three-fourths of respondents have consistently thought nuclear power will play a very important or somewhat important role in meeting the nation's future electricity needs. This finding is consistent with public knowledge of the importance of nuclear-generated electricity and with the public's pragmatic acceptance of the status quo. The picture changes somewhat when Americans are asked whether the nation should rely more on nuclear power in the future. In two polls, respondents were slightly more inclined to opt for less nuclear power in the future.

6.3 Nuclear as a Trade-Off Choice

Nearly all imported oil is used for petroleum products and not for the generation of electricity. Nevertheless, Americans have been asked whether they would favor nuclear power if it would reduce the nation's dependence on foreign oil. Presented with such a trade-off, majorities of the public favor nuclear power—often by sizable margins.

The modern trade-off context is the question about substituting nuclear for fossil fuels (especially coal) in the generation of electricity to reduce greenhouse gas emissions. What is the level of public support for nuclear power to address global warming? Data directly addressing this question are meager. Surveys from the 1990s show that majorities believed that nuclear energy could cut greenhouse gas emissions and air pollution and vast majorities would favor using nuclear energy to cut greenhouse gases and air pollution. Recent data, however, contradict that finding. Asked whether they would favor nuclear power as a means “for dealing with the pollution that causes climate change,” a majority of citizens (55 vs 42%) were opposed. Thus, we have a far from complete answer to this question.

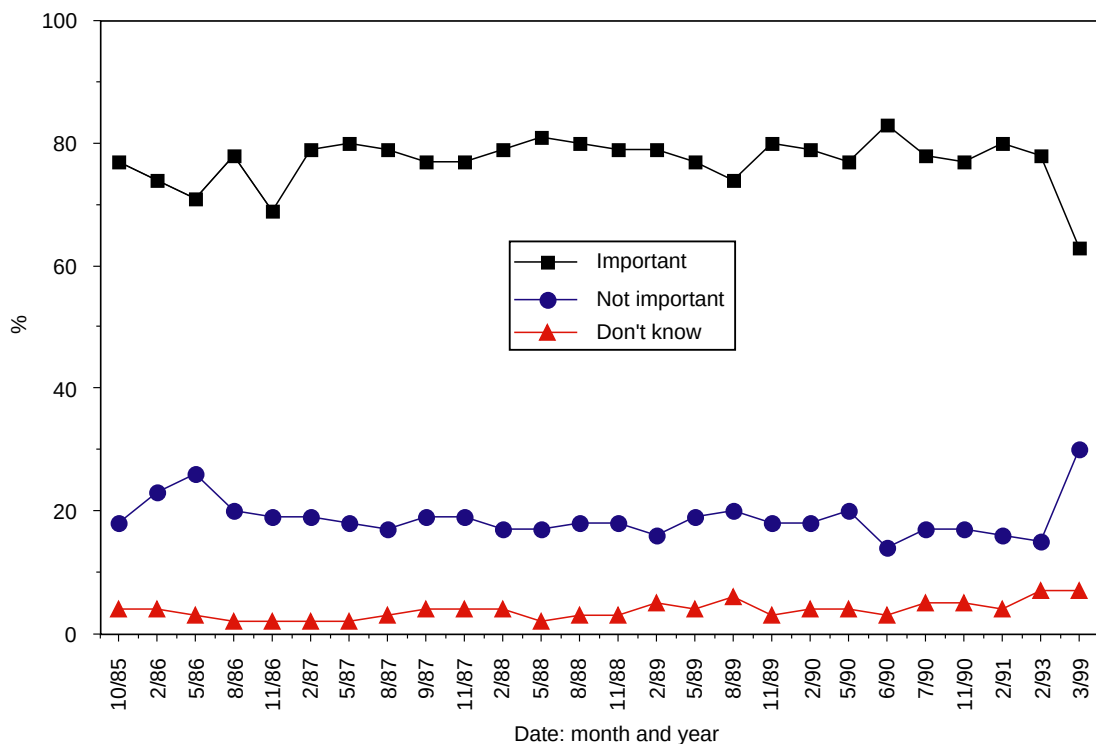


FIGURE 5 Importance of nuclear power in the future. Question: “How important do you think nuclear power plants will be in meeting this nation’s electricity needs in the years ahead?” Source: Cambridge reports and Gallup national samples.

6.4 Interpreting the Public Mood

Taken together, the overall results may appear contradictory, perhaps even irrational. On the one hand, solid majorities of the American public oppose the construction of more nuclear power plants and are likewise opposed to their local siting. On the other hand, equally solid majorities believe nuclear power will be an important energy source in the nation's future, although this is not necessarily their first choice. What accounts for the apparent contradiction? In our view, the results do not reveal a contradiction but instead a pragmatic logic whereby the public distinguishes between nuclear power in principle and in practice.

For nuclear power, it appears that the collective public's preference for specific policy actions can be quite different from its general attitudes. Thus, Americans appear generally to support a role for nuclear power in the future and are solidly behind the idea of leaving the nuclear option open—perhaps as a trump card against possible future energy shortages. However, when it comes to the specific instrumental means for achieving that future—namely, the siting and construction of nuclear power plants—they are solidly opposed. What are the key impediments to the realization of the public's expected nuclear future?

7. NUCLEAR IMAGERY

Underlying attitudinal responses and global assessments of nuclear energy are a variety of cognitive filters as well as images that shape thoughts and dispositions toward the technology. Although nuclear energy is Janus-like, with inherent benefits and risks, there is considerable evidence showing negative imagery is much deeper than positive imagery, overpowering the positive with salient thoughts of danger, toxicity, death, sickness, destruction, pollution, and the like. Some researchers have gone further, arguing that nuclear facilities, especially nuclear waste facilities, attract stigma—a deeply discrediting or devaluing feature that marks them for special concern or avoidance. Although the claim for stigma may stretch the available evidence, it is unequivocally clear that images associated with things nuclear are predominately negative and militate against the unreserved acceptance of the technology. There is little doubt that the future of nuclear energy is heavily dependent on overcoming this deep negative imagery.

8. NUCLEAR WASTE

One of the unavoidable by-products of nuclear-generated electricity is the production of a variety of highly radioactive wastes. Originally believed to be a tractable, and solvable problem, a solution to the disposition of nuclear waste still eludes the nuclear establishment worldwide. Although there is a universal scientific consensus that deep geological isolation is the preferred solution to the disposition of wastes—especially high-level wastes with their extremely long radioactive half-lives—no nuclear nation has built a geologic repository or has emplaced any waste. With few exceptions (e.g., Finland), no nation has even been able to achieve a sufficient level of political and public support to reach an uncontested final decision on the siting of a repository. In the meantime, the waste produced by nuclear energy accumulates in pools at reactor sites or at military installations.

A failure to solve the waste problem proves to be especially challenging in gaining public acceptance of the technology. Publics everywhere hold deep negative images of nuclear waste—even deeper than for nuclear power—and are opposed to siting a repository not only in their immediate vicinity but also just about anywhere. When asked to indicate the closest distance they would be willing to live in the vicinity of a variety of facilities, a nuclear waste facility typically ranks last—in one survey at 200 miles away (twice the distance from the next most undesirable facility, a chemical waste landfill, and three to eight times the distances from oil refineries, nuclear power plants, and pesticide manufacturing plants).

Particularly revealing of the depth of negative imagery associated with nuclear waste are results from a survey in which citizens were asked to associate freely about the concept of a nuclear waste repository. The vast majority of images elicited were extremely negative, punctuated by such categories as dangerous/toxic, death/sickness, environmental damage, and bad/negative.

Thus, a failure to solve the nuclear waste problem could be the Achilles' heel of nuclear power, ensuring a continuation of the technology's cloudy future or worse—its ultimate failure. That the ultimate success of nuclear power is conditioned by a solution to the waste problem is underscored in a 2001 Eurobarometer poll conducted in 15 European Union (EU) countries. When asked, "If all waste is managed safely, should nuclear power remain an option for electricity production in the EU?" 51% of the 16,000 respondents answered in the affirmative, whereas

near equal percentages rejected this option (25%) or had no opinion (24%). This conditional favorability, however, is tempered by the strong plurality (46%) of the same respondents who said they thought that there was no safe way to dispose of the waste.

9. CONCLUSION

If it were written in ancient times, the history of nuclear power in the United States and elsewhere would be written as a drama between two mythical figures: the Phoenix and Hydra. Like the mythical Phoenix that arose from its own ashes, the nuclear establishment—a combination of scientific, commercial, and policy interests—counters each setback to nuclear's fortunes with either organizational or technical fixes. The emergence of a fourth generation of reactors, presumably much safer than their predecessors, is the most recent manifestation of this pattern. In contrast, since the mid-1970s the general public has viewed nuclear power as a many-headed Hydra in which as soon as one of its concerns is addressed—or recedes from consciousness—two others spring up to take its place.

Since the first sustained chain reaction at Stagg Field at the University of Chicago in 1942, commercial nuclear power has experienced paroxysms of enthusiasm and growth, followed by extended periods of reflection and reconsideration. Public support for the technology began with unrestrained enthusiasm, only to be followed by extended periods of doubt and reservation. Currently, two clouds of uncertainty continue to shroud public acceptance of the technology: the safety of nuclear plants and the challenge of disposing of nuclear wastes, especially high-level wastes.

Whether public acceptance can be regained with a new generation of reactors with advanced designs—argued to be inherently safer than the current generation of reactors—will depend on a sequence of contingencies. It will depend on whether the claims about safety are credible, whether credible arguments about safety can overcome the deep negative imagery that shrouds the technology, whether alternative technologies emerge and are economically competitive, and on the vagaries of the future geopolitics of energy supplies. It will also clearly depend on solving the nuclear waste problem.

Irrespective of the regaining of public trust for nuclear power and its management, it is unlikely that the technology will be abandoned worldwide—

at least not in the foreseeable future. Too much has already been invested in nuclear infrastructure (including many national and international institutions), too many national states see a significant role for nuclear in their future energy mix, and too many powerful technological elites remain committed to promoting its growth. However, many questions remain. Will nuclear power remain in steady state (eventually resulting in its demise as the current inventory of plants reach the end of their usable lives) or grow again? Will it be able to grow absent a successful program for the disposal of nuclear wastes? If it does grow, will it be in a supportive public climate or a climate of continuing tension, forcing the government to take a more active role in overcoming or circumventing those tensions—perhaps with legislative or regulatory actions that erode the democratic process? All these questions embed even deeper and more challenging ones about technological knowledge, historical contingency, political culture and climate, and the evolution of democracies. It will be the answers to these hoary questions that will determine the future of nuclear power.

SEE ALSO THE FOLLOWING ARTICLES

Environmental Injustices of Energy Facilities • Nuclear Engineering • Nuclear Power: Risk Analysis • Occupational Health Risks in Nuclear Power • Public Reaction to Electricity Transmission Lines • Public Reaction to Offshore Oil • Public Reaction to Renewable Energy Sources and Systems

Further Reading

- Bupp, I. C., and Derian, J.-C. (1981). "The Failed Promise of Nuclear Power: The Story of Light Water." Basic Books, New York.
- Dunlap, R. E., Kraft, M. E., and Rosa, E. A. (eds.). (1993). "The Public and Nuclear Waste: Citizens' Views of Repository Siting." Duke Univ. Press, Durham, NC.
- Freudenburg, W. R., and Rosa, E. A. (eds.). (1984). "Public Reactions to Nuclear Power: Are There Critical Masses?" Westview /AAAS.
- Garwin, R. L., and Charpak, G. (2001). "Megawatts and Megatons: A Turning Point in the Nuclear Age." Knopf, New York.
- Jacob, G. (1990). "Site Unseen: The Politics of Siting a Nuclear Waste Repository." Univ. of Pittsburgh Press, Pittsburgh, PA.
- Jasper, J. (1990). "Nuclear Politics: Energy and the State in the United States, Sweden, and France." Princeton Univ. Press, Princeton, NJ.

Lovins, A. (1976). Energy strategy: The road not taken. *Foreign Affairs* 55, 65–96.

Rosa, E. A., and Dunlap, R. E. (1994). Nuclear power: Three decades of public opinion. *Public Opin. Q.* 58, 295–325.

Schwartz, S. I. (1998). “Atomic Audit.” Brookings Institution Press, Washington, DC.

Weart, S. (1988). “Nuclear Fear: A History of Images.” Harvard Univ. Press, Cambridge, MA.

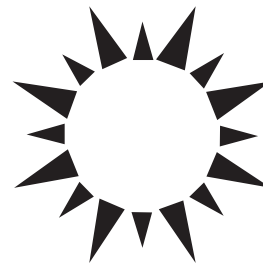
Public Reaction to Offshore Oil

WILLIAM R. FREUDENBURG

University of California, Santa Barbara
Santa Barbara, California, United States

ROBERT GRAMLING

University of Louisiana, Lafayette
Lafayette, Louisiana, United States



-
1. Introduction
 2. Historical Overview
 3. The U.S. Tidelands Controversies
 4. The International Growth of Offshore Technologies
 5. The Evolving Range of Views Throughout the World
 6. Factors Associated with Differing Viewpoints in the United States

Glossary

attitudes Unlike physical objects, attitudes cannot be seen or weighed. Instead, they are more like the biological concept of life—difficult to define but sufficiently important and pervasive to have attracted considerable study. In general, they are seen as being like opinions but more deeply connected to one's overall orientation toward the world (and hence less likely to change) while not as deeply connected as values. In most of the technical literature, attitudes are seen as having three components—relatively factual beliefs, more emotional evaluations, and in many but not all definitions, “conative” or behavioral orientations, involving intentions to act in certain ways but not others.

Continental Shelf The relatively shallow ocean-bottom lands close to the edge of a continent. In many but by no means all cases, these lands do form a literal shelf, sometimes extending for dozens or even hundreds of kilometers beyond dry land before dropping off into deeper waters. In the United States, the outer Continental Shelf includes the ocean-bottom lands beyond the limits of state jurisdiction (generally 3 miles from land).

Dutch disease A term that is sometimes used to describe almost any case of economic weakness in the face of resource exports but that more formally refers to economic weaknesses resulting from twin problems experienced in The Netherlands in the wake of its North Sea oil discoveries: (i) a significant increase in the strength of national currency (making its other exports

more expensive) and (ii) a tendency for the oil sector to draw people and capital away from other sectors of the economy, such as manufacturing and agriculture.

economic multiplier Many approaches to socioeconomic impact assessment divide local or even national economies into two broad sectors: economic base (or export) activities, which bring cash into an area through the selling of goods and services to people who live elsewhere (e.g., agriculture, oil extraction, and tourism), and nonbasic (or service) activities, which can create additional jobs by keeping the cash circulating locally (e.g., when people use that cash while visiting local shoe repair shops, restaurants, grocery stores, etc.). The ratio between basic and nonbasic activities provides the economic multiplier, although actual multipliers from new activities (e.g., new oil drilling) have often proven to be much lower in practice, in recent decades, than the ones that have been put forth in the proponents' economic models.

social multiplier Just as one economic base job can help to support other nonbasic jobs in the same community or region, one person who works for a given industry is likely to have many friends and relatives nearby, particularly in the case of an established industry. Since people tend to care not just about their own well-being but also about the well-being of their friends and acquaintances, there can be much stronger political support for an important industry (e.g., agriculture in Iowa and oil in Texas) than might be expected simply on the basis of the number of people or jobs supported by that industry directly.

tidelands controversies A series of battles during the mid-20th century involving political control over (and tax revenues from) new activities of offshore oil drilling in U.S. waters. The controversies were largely brought to an end by two pieces of legislation, one of which gave control and revenues to specific states for drilling taking place in territorial waters (generally those within 3 miles of shore) and the other of which granted control and revenues to the federal government for activities farther offshore.

It was more than a century ago when humans first thought of the idea of drilling through water, as well as land, in search of oil, and during the course of this century, the attitudes or views about potential oil developments have varied widely. Rather than representing relatively random reactions, or anything approaching universal approval or opposition, local attitudes tend to vary more systematically across time and space. These attitudes, in turn, shape the policies and policy actions that can cause offshore oil development to be easier or more difficult to pursue. This article discusses the historical background and geographic extent of offshore oil development and the importance of three factors with regard to a region's attitudes toward offshore development: the region's historical relationship to oil development; the biophysical environment of the region, such as its coastal and marine topography; and relevant socio-cultural conditions, such as educational levels, social networks, and prior economic adaptation.

1. INTRODUCTION

The first offshore drilling was actually done from a set of oil derricks along a pier that jutted into the Pacific Ocean off the shores of Summerland, California, southeast of Santa Barbara, in 1898, and the first offshore oil controversy erupted a few miles away the very next year. When an oil company began to construct an oil derrick off the shores of Montecito, California—the highly affluent Santa Barbara suburb that is adjacent to Summerland and that occupies the few miles of coastline between Summerland and Santa Barbara—a local mob, described approvingly on page 1 of the *Santa Barbara Morning Press* the next day as “a party of the best known society men of Santa Barbara, armed to meet any resistance,” attacked the rig and tore it down.

The local “society men” seem not to have suffered any noteworthy legal repercussions from their actions, despite having been so well-known, but oil companies did: The graphic expression of local attitudes proved effective in blocking further drilling along that stretch of coastline for decades to come. Elsewhere in the world, however, for at least the next 50 years, most such efforts to find oil—offshore as well as on—were much more likely to be the focus of celebration, not protest. Had this article been written in 1953 rather than in 2003, it almost certainly would have included a discussion of what were known at the time as the “tidelands controversies,”

but as noted later, these controversies had to do not with opposition to offshore oil but with the virtual opposite, involving competition over which political jurisdictions would be able to enjoy the accompanying tax benefits.

Today, offshore exploration and production have spread throughout the world, taking place off the coasts of Africa, Australia, Central and South America, and Europe. In the United States, offshore oil is limited to a handful of rigs off Alaska's North Slope, 21 rigs off California (all in the Santa Barbara region, just a few miles away from Summerland and Montecito), and a vastly larger number of less famous oil rigs (approximately 3500) in the Gulf of Mexico, where some platforms are producing oil in water that is 2 miles deep.

As these patterns suggest, public attitudes toward offshore oil activities are far more diverse than once might have appeared to be the case. To understand what has happened in the century or more since the derricks first appeared on the Summerland pier, and to understand the variations in public attitudes toward offshore oil activities, it is useful to begin with a brief history. Following the historical summary, we present an overview of the variations that can be found in some of the key offshore regions of the world and then provide a more detailed examination of what appear to be the two polar cases—the two regions, both in the same country, where attitudes appear to be the most and the least favorable.

2. HISTORICAL OVERVIEW

The original impetus behind the search for offshore oil came just 17 months after the displeasure of the Montecito mob, when the world's first major petroleum reservoir, Spindeltop, was tapped on January 10, 1901, near Beaumont, Texas. Compared to the petroleum reservoirs that had been found in Pennsylvania and elsewhere during the latter part of the 19th century, Spindeltop was huge; its discovery transformed the petroleum industry, both in the United States and throughout the world. For purposes of this article, however, the key point is that the reservoir proved to be located under the limestone cap that covered the top of a salt dome. Not surprisingly, this discovery led to interest in other such domes, many of which are located off the shores of Louisiana and Texas.

Despite the early discoveries off the shores of south-central California—or perhaps in part because

of the reactions they inspired—it was not until the 1920s that offshore oil development began to be pursued in earnest. The pursuit took place largely under the waters of Lake Maraicaibo in Venezuela and more extensively in the marshes and the shallow, protected waters of the Gulf of Mexico. By far the greatest development took place along the Louisiana deltaic plain—the lands that were formed by the delta of the Mississippi River—once the development of nondestructive exploratory (seismic) techniques allowed oil companies to locate and explore the numerous salt domes along the Louisiana and Texas coasts of the Gulf of Mexico.

At the time, it may not have occurred to many of the oil companies that explored the Louisiana marshes, or the people who lived nearby, that they were involved in the early stages of developing what would become one of the great industrial innovations of the 20th century. What almost certainly did occur to them was that it was no simple matter to drill oil wells through the unstable soils of the region, especially in that the best oil prospects were often situated in marshes, swamps, and the shallow waters of bays and lakes. The initial techniques involved the construction of a platform on pilings that were driven into the coastal marshes, with drilling equipment brought to the site on barges and installed on the platform. From there, drilling proceeded much as it would have been done on land, except that transportation to and from the site was often by boat, and additional costs were created by the need to build a platform for each exploratory well. In addition, pilings driven into the thick layers of silt have no solid foundation, relying entirely on the friction between the pilings and the mucky bottom. Since drilling machinery produces significant vibrations, up to 250 pilings were needed for a single exploratory platform. As interest in deeper prospects increased, these early techniques became too costly, since the nonrecoverable cost of platform construction was a constant regardless of the outcome of the drilling.

The technological breakthrough that facilitated large-scale exploration came in 1933 when the Texas Company (later Texaco) introduced the “submersible” drilling barge: A drilling rig was simply mounted on a barge, which could be towed to a drilling site and flooded. Sitting on the shallow bottom, the barge provided a stable base for drilling; once the job was done, the barge could be refloated and towed to a new location. Lake Pelto, Louisiana, behind Last Island, east of the mouth of the Atchafalaya River, was the site of the first successful test of the technology, opening a new era in marine drilling. Within a

decade, these rigs were in use throughout the marshes and shallow coastal waters in Louisiana and Texas, sparking a growing support sector for marine drilling.

Throughout the 1930s and 1940s, the submersible drilling rigs evolved, becoming larger and more powerful, allowing them to tap the deeper oil-bearing formations along the Gulf Coast. Steam engines were replaced by diesel-electric rigs, and the tungsten carbide drilling bit was introduced in 1952. By 1955, inland submersibles were setting world drilling depth records of 22,500 ft. Much of the initial exploration, drilling, and production occurred in eastern St. Mary Parish, Louisiana, and residents of nearby towns, primarily Morgan City and Berwick, appear to have been eager to help provide labor and support to the oil firms that were drilling in the challenging areas nearby.

As inland drilling rigs were evolving, the focus was shifting offshore. In 1946, the Magnolia Petroleum Company (which later became part of Mobil), using pilings driven into the sea bottom, constructed a platform in the open Gulf on a site leased from the state of Louisiana. The site was south of Morgan City, 5 miles from the nearest land. Although the attempt was an economic failure (a dry hole), it was the key forerunner of what most observers today would consider to be a truly offshore effort, and it was a technological success in that it demonstrated that exploration in the open Gulf was feasible.

Following Magnolia's lead, and using the same type of technology, Kerr-McGee established the first producing well on the outer Continental Shelf in 1947, 12 miles off the Louisiana coast, almost due south of the mouth of the Atchafalaya River, on what is now Ship Shoal Block 32. This sparked additional interest offshore, and by the middle of 1948, in addition to offshore production by Kerr-McGee, 13 locations were in various stages of development off the coasts of Louisiana and Texas and plans were under way for drilling in 14 other locations. Within the next decade and a half, production platforms would be found in more than 200 ft of water, although the offshore bottoms of the Gulf of Mexico were sufficiently gradual that these platforms were being placed many miles offshore.

3. THE U.S. TIDELANDS CONTROVERSIES

The virtual explosion of offshore drilling activities during this era provided the backdrop for what came to be known during the Truman administration and

afterwards as the Tidelands controversies. These controversies, which continued until the subsequent Eisenhower administration pushed two new laws through Congress, were focused squarely on offshore oil drilling, but contrary to what many present-day readers might expect, they had almost nothing to do with public opposition. Instead, they had to do with what political body would reap the benefits of offshore oil—the states or the federal government.

In addition to the oil-friendly states of Texas and Louisiana, which were engaged in legal battles with the federal government, California and Florida—despite the intense opposition that both states would later express toward offshore oil and that at least some areas of California had shown a half century before—also showed strong interest in the offshore oil industry. As one observer informally summarized the tenor of that earlier time, looking back from a perspective of three decades later, “The basic attitude was, ‘Good luck—hope you find some oil—and if you do, send us a check.’” The main concerns of the time had to do with whether the regulations and the public revenues from offshore oil would be the business of the federal government or the states.

Contrary to the Truman administration, which had inspired the controversies by favoring federal control, the Eisenhower administration espoused a “states’ rights” position. This position aligned well with the preferences of the oil companies (which expected more favorable regulatory policies from the states than from the federal government) and with the relevant states (which were mainly located along the Gulf of Mexico and which were keenly interested in the public revenues that oil development could generate). The Eisenhower administration’s first piece of legislation thus inspired considerable support from the relevant states and the oil companies by giving the states the legal and financial rights to all oil developments taking place within 3 miles of their shores. The second piece of legislation, which appears to have been the focus of much less attention at the time, gave the federal government the rights to the undersea lands that were farther away from the continent, along what came to be called the outer Continental Shelf, or the subsea lands that were out beyond the 3-mile limit of the states’ jurisdictions.

4. THE INTERNATIONAL GROWTH OF OFFSHORE TECHNOLOGIES

Throughout the 1960s and early 1970s, most of the technology and the expertise still came from the

Louisiana and Texas shores of the Gulf of Mexico. In 1968, for example, the *Glomar Challenger*, a state-of-the-art drill ship from Orange, Texas, began a series of worldwide trips for the Scripps Institute, taking deep-bottom cores for research purposes and using technology that would have been inconceivable only a few years before. Before the trips were over, the *Glomar Challenger* would drill a hole almost 3000 ft deep in an ocean-bottom location that was more than 16,000 ft below the surface of the sea.

As exploration and production continued to expand, however, offshore energy development came to be an increasingly worldwide phenomenon. By the mid-1960s, exploratory drilling had begun in the Persian Gulf, along the western shores of Africa (especially off Nigeria), in the Far East (Indonesia, Japan, Borneo, and Australia), and along the Cook Inlet in Alaska. By 1967, 75 countries were exploring for offshore oil and gas, and 20 of these were producing oil and/or gas offshore.

It was the North Sea experience, in particular, that may have provided the clearest early warnings of what would follow. Interest in offshore petrochemical extraction from the North Sea had begun in the late 1950s, just a few years after the passing of the Eisenhower-era legislation in the United States, and growing significantly in 1959 with the discovery of the Groningen gas field off Holland. By 1962, when the size of the Groningen field was better understood, that region saw a multimillion-dollar drilling effort, which was in full swing by the mid-1960s, with the first North Sea production in 1967. As offshore drilling efforts continued to expand, however, the offshore industry found itself in waters that were literally as well as figuratively deeper and more hostile. The loss of the drilling rig *Sea Gem* in the North Sea in 1965 indicated the dangers associated with the new environment, and increased emphasis was placed on adapting the Gulf of Mexico technology to more hostile conditions.

Particularly after the 1969 Santa Barbara oil spill just a few miles away from the first offshore oil derricks—in a location where the Nixon administration had insisted on proceeding with oil production despite local worries about just the kind of spill that actually occurred—the major oil companies devoted increasing attention to other areas of the world, reinforcing the already aggressive efforts by drilling and support companies that were searching for offshore discoveries. By the mid-1970s, for the first time since the earliest days of offshore drilling,

the majority of the world's offshore drilling operations were located in countries outside the United States, with approximately 75% of the activity in other countries, and with the North Sea being the prime area.

As previously mentioned, however, in testing the waters of the North Sea, oil companies encountered much colder temperatures than they had experienced in the Gulf of Mexico, politically as well as climatically. Rather than being able simply to "send a check" after finding oil, the oil companies (which in the case of Norway included a state-owned oil company) needed to spend years in negotiations with the citizens and leaders of the region. In both of the nations exercising political jurisdiction over key areas of the North Sea—Norway and Great Britain—affected citizens expressed a range of concerns and demonstrated a level of political resolve that would have been almost unimaginable along the Gulf of Mexico in earlier times. In a nation that takes great pride in its long history of Scandinavian distinctiveness, its widespread public participation, and its careful planning for environmental preservation, Norway understandably provided a very different context for operations from that of the oil industry's previous experiences in the Gulf of Mexico. Similarly, in Great Britain (particularly in Scotland and the Shetland and Orkney Islands, which would have had the most to gain or lose from offshore development), there were extensive discussions of the kinds of conditions under which the development of North Sea oil would be allowed to proceed.

The net results of the extensive discussions in both Norway and Great Britain included numerous measures for environmental protection, for local as well as national economic benefits, and for local socioeconomic impact mitigation, which were seen as innovative or even revolutionary by many analysts, especially against the backdrop of previous U.S. experiences. Norway, in particular, emphasized the combining of environmental protection with economic development measures (including a state-owned oil company, but also a number of private enterprises) that eventually led the country to become one of the world's leaders in offshore technology, allowing it to emulate the Gulf Coast pattern of selling not only its oil but also its technology. Several years later, when development began on the giant Hibernia oil field off the coast of Newfoundland, Canada, most observers saw far more similarity to the North Sea model than to previous experiences along the Gulf of Mexico.

5. THE EVOLVING RANGE OF VIEWS THROUGHOUT THE WORLD

There is no single or universal attitude toward offshore development. Instead, as can be seen by contrasting the regions that represent the two most extreme ends of the continuum—a task to which we will soon turn—the attitudes that are encountered are likely to reflect the specific physical, historical, and socioeconomic conditions that are in effect for any given place and time. To the extent to which there are shared characteristics across locations, most of the main reasons for supporting oil developments in a given locality are economic ones, and most of the reasons for opposing the same developments are environmental ones. Even this generalization, however, is oversimplified. As discussed later, many of the regions that have hosted oil development have had fewer economic benefits than they expected. However, at least in the decades since the Santa Barbara oil spill of 1969, the most often feared environmental problem of offshore drilling, the oil spill, has proved to be quite rare—at least in offshore drilling and production, as opposed to the far more numerous spills associated with the transportation of oil, particularly the well-known tanker spills from the Exxon Valdez, the Braer, and other such ships. Particularly in industrialized nations, the environmental and economic challenges and problems that are of greatest concern to nearby residents may have little to do with oil spills, having to do instead with the onshore activities that support the offshore platforms. Some of these activities and businesses, including those involving diesel mechanics, suppliers of drilling muds, and the many tons of pipes and other supplies, can conflict with other ways of using prized coastal locations, many of which are also quite important economically as well as esthetically, with uses ranging from fishing to high-value homebuilding.

It is important to be particularly cautious, however, about drawing many firm conclusions about public attitudes in less developed countries, because systematic research on attitudes in most such countries is relatively rare. Far from displaying high levels of concern or sensitivity toward local attitudes, in fact, many of the less prosperous nations that have developed their oil resources have shown tendencies toward political repression that many observers have found disturbing. So pronounced is this tendency—onshore as well as offshore—that after examining cross-national data from 113 nations, Michael Ross published his findings under the title "Does Oil Hurt

Democracy?" The answer, he concluded, appears to be in the affirmative.

In broad terms, to the extent to which the reactions of many different nations have been examined, the most common focus has involved the financial and, to a lesser extent, the political implications of offshore development. As a useful simplification, three main patterns have emerged across time. The first and by far the most common pattern in some respects reflects a combination of the earlier tidelands controversies in the United States and the efforts by Norwegian leaders to use planning and politics to extract greater economic benefits from the vast flows of revenue that are involved in offshore oil development. The second involves an underlying pattern of concern regarding the kinds of democracy-harming tendencies noted by Ross. The third pattern provides an ironic contrast against the first, involving a problem that is sometimes called Dutch disease, as will be discussed later.

In Venezuela, the discovery of oil under the waters of Lake Maraicaibo in the early 1920s was followed by more than a half century of economic growth, during which oil revenues were used to provide a wide range of economic development benefits. For most of the next 50 years or more, Venezuela was seen by most observers as one of Latin America's most notable models of democratic stability and social peace, and the common conclusion was that the political as well as the economic benefits could be traced in large part to oil-based prosperity. Since the oil price collapses of the 1980s, however, Venezuela's political stability has been seen by some to be as questionable as its prosperity, with clear signs of tension between the middle/upper classes and Hugo Chavez (the current President at the time of this writing in 2003). Recent assessments are much more likely to call attention to class warfare than to social stability.

A similar pattern is visible in Indonesia, which provided the lion's share of "Dutch" oil for Royal Dutch Shell for nearly half a century, during which time the former colony was generally known as the Dutch East Indies, before achieving independence after World War II. For a period of several decades after the former colony became an independent nation, its leaders were seen by most observers as showing considerable political acumen in their efforts to obtain economic and developmental benefits from indigenous oil resources. Rather than attempting to "nationalize" or take over the oil companies that were operating in Indonesia at the time, Ibnu Sutowo—a close personal associate of Presidents Sukarno and Suharto and the key force

behind Indonesia's state-owned oil company, Pertamina—emphasized a more gradual approach. One component involved the motto "Learn while you work, work while you learn" for the Indonesian nationals who continued to work in the oil fields until the old patterns of colonial dominance (and the old colonial Mining Law of 1899) were finally brought to an end. Another, more tangible component involved the relatively gradual process of ending the earlier patterns, beginning when Article 33 of the new Constitution of 1945 declared all of Indonesia's natural resources to belong to its people but not fully completed until approximately 20 years later, when the Petroleum Act of 1960 and then the production sharing contract arrangement of 1966 ultimately came into effect.

Even after Indonesia passed its new law in 1960, the major oil companies were not taken over by the Indonesian state; instead, the companies experienced a transition from a previous legal status of "concessionaire" (i.e., enjoying concessions that had been granted under Colonial Dutch rule) to a new status of being contractors to the state enterprise, Pertamina. The transition was completed after Sutowo and the Indonesian state instituted production sharing contracts, which required foreign contractors to bear the full cost of exploring for oil but reserved approximately 40% of the oil for Pertamina and the Indonesian government. Under other contract provisions, most of the lands that were leased to the oil companies reverted back to the government within a decade or less, and the Indonesian share of the profits would increase as world oil prices increased.

Not surprisingly, it appears that at least the financial implications of Indonesia's increasing control over its oil and the associated revenues won considerable praise from most Indonesians for many decades. Indonesia has also been seen by many scholars as an exception to some of the threats to democratic development that are of concern to authors such as Ross. The oil revenues were reinvested in a wide range of other enterprises that were seen as contributing both to the development of the growing national economy and to the integration of what is not only the most populous petroleum-exporting nation but also the most heavily populated Islamic nation on Earth. This latter point is a major one, given that Indonesia embraces what most Western observers find to be a surprisingly diverse and far-flung collection of peoples. It encompasses more than 1000 islands and hundreds of native languages spread across a distance of more than

3000 miles of land and water. Through investments in everything from schools to steel factories, Indonesia used its oil revenues as an important resource for building a stronger and more prosperous nation.

Particularly in the later years of the Suharto era, however, Indonesians became increasingly critical of the “crony capitalism” and corruption that appear to have become increasingly pervasive, both in Pertamina and in the broader society, and of the increasing extent to which the income derived from offshore oil and gas fields was used to support military repression. As suggested by the fact that some of the more troubled areas of Indonesia also happen to have rich offshore petroleum deposits—including Aceh, East Kalimantan, and the now-independent nation of East Timor—the degree of Indonesian “exceptionalism” may be less than previously thought.

Finally, off the western shores of Nigeria, very few observers have seen the kinds of exceptionalism once thought to exist in Indonesia and Venezuela. Instead, particularly after what the Nigerian government characterized as the “execution” of the noted human rights activist, Ken Saro-Wiwa, there has been extensive international attention to what many have seen as heavy-handed and antidemocratic central government repression of local opposition to offshore oil development and its attendant onshore impacts, particularly among the native people of the area, the Ogoni.

Economic benefits have sometimes proved elusive, as well. Although the vast sums of money involved in offshore oil development often lead to expectations for rapid economic growth in oil-exporting countries, careful analyses have tended to find much less desirable outcomes, economically as well as politically.

Perhaps the best known description of the economic problems is the notion of Dutch disease. Given its catchy alliteration, this phrase has been attached to almost any example of economic weakness in the face of resource exports, but more formally the phrase is understood to involve the combination of two factors that appear to have led to a generally unexpected slowdown in the economy of The Netherlands, even in the face of that country’s increased revenues from its own North Sea petroleum reserves. First, a sharp increase in exports led to a significant increase in the strength of the national currency—a fact that meant other exports from the country became more expensive for people who lived outside The Netherlands, reducing international demand for Dutch products. Second, the booming resource sector was seen as drawing people and

capital away from other sectors of the economy, such as manufacturing and agriculture, raising the effective production costs in the other sectors over time. These concerns led to deliberate efforts, particularly in Norway, to offer increased inducements for investment in non-oil sectors—efforts that may well have been warranted in any case. However, subsequent analyses have indicated that unexpected combinations of rich resources and poor people appear to be more widespread than can be explained by the Dutch disease hypothesis alone.

6. FACTORS ASSOCIATED WITH DIFFERING VIEWPOINTS IN THE UNITED STATES

There is no one or universal pattern of public attitudes toward offshore oil development. Instead, attitudes vary widely in response to local conditions and the nature of oil development that is proposed. Of all the places that have been studied systematically, however, it appears that the widest variations in attitudes are found between two different areas of the same country—northern California and southern Louisiana. It is clearly not the case that the opposition is a straightforward function of the environmental damage that has been created in each region. Indeed, with the possible exception of the Santa Barbara oil spill (which took place along California’s south-central coastline, in any event), the environmental impacts of offshore oil development have actually been far more significant and widespread in southern Louisiana than in northern California. Instead, the variations in public attitudes between these two regions tend to illustrate the broader points we are making in this article, reflecting the historical, physical, and social characteristics of the regions in question. Thus, a comparison of the two regions provides a useful way to bring this article to a close.

Historical factors are important both because of changes in broader historical factors and local cultural characteristics, and because differing regions experience an activity (in this case, offshore oil) during different periods of that historical transition, in effect resulting in different local histories. When oil exploration started in Louisiana’s coastal marshes in the 1920s, the prevailing attitude toward the environment almost everywhere in the United States was an exuberant one, involving the “conquest of nature” and the fresh legacy that remained from the mythic

conquest of the western frontier. Although environmental concern was not completely unknown, it certainly had little of the character or strength that it would have a few decades later. In nearby Florida, for example—where later decades would see a level of organized opposition to offshore oil drilling comparable to that of California—essentially all of the offshore lands that were seen as potential oil-bearing prospects were leased to oil companies by the state of Florida well before anyone there had ever heard of Earth day.

In addition however, important aspects of the development of offshore technology were unique to southern Louisiana. When oil exploration began to move into the coastal marshes, the marshes were viewed not as ecologically vital wetlands but as hostile territory that needed to be subdued for human benefit. Louisiana also had a long history, dating back to the early 1800s, of resource harvest (fish, shrimp, oysters, and fur cypress lumber) in coastal areas. Furthermore, oil and these earlier extractive activities, especially fishing, proved to be highly compatible in several important ways. First, oil companies had no experience with coastal and offshore waters, but the fishing industry did. When Magnolia Petroleum Company drilled the first offshore well in 1946 south of Morgan City, boats from the local shrimp fleet were hired for the initial survey and the transportation of workers and supplies. Local people supplied the navigation knowledge for early offshore oil, and some of them later became crew and supply boat captains.

Second, the scheduling of offshore oil work allowed for individuals to maintain traditional coastal occupations. The movement into the marsh and estuaries of the Louisiana deltaic plain was large scale and relatively permanent. However, since it involved environments in which human settlement was not really possible, and in which distances and transportation difficulties meant that commuting had to be limited, an innovative form of concentrated work scheduling soon evolved. The common pattern for offshore work became one in which employees meet at a prearranged site to “go offshore” either by boat or, recently, by helicopter. The stay offshore varies, but is typically 7, 14, or 21 days. Following the stay offshore, the employees return to the meeting site and have a period of time off, typically the same length as the stay offshore. Because of this schedule, many such workers chose to continue their traditional occupations. This meant that boats, gear, and knowledge were maintained and did not fall into disrepair, as might have occurred if oil jobs in

Southern Louisiana had been scheduled for five days a week. When the oil patch crashed in the mid-1980s, many coastal residents found it useful to fall back on these skills and equipment to feed their families.

In addition, oil development in Louisiana started gradually, and it grew as the product of local innovation, which understandably generated considerable pride. In contrast, when the first offshore development appeared in California in the late 1960s, not only was environmentalism emerging as a national trend but also the huge rigs that were by that time associated with offshore oil were seen by a largely urban population as foreign and threatening. When the Nixon administration insisted on moving ahead with offshore leasing just a few miles away from the beaches of Summerland and Montecito in the face of stiff local opposition, and after the Santa Barbara oil spill a few years later provided a tangible indicator of the validity of local concerns, Californians’ opposition to further offshore oil activities became all the more intense.

Physical factors, or to use the more technical terminology, coastal geomorphology, can also be seen to differ quite distinctly between the two states. There are fundamental differences between the Louisiana and California coasts, both in coastal topography and in marine topography. The coastal regions of Louisiana differ from those of California, as well as from most other areas of the United States, in three main ways: Coastal wetlands limit land-based access to the coast, estuaries offer many opportunities for harbor space, and low relief and low energy levels characterize most of the coast. As a result, the distribution of populations and roadways in Louisiana is very different from that found in most coastal states. Few of the state’s population lives on or near the coast and, with the exception of Grand Isle, it is often almost impossible to get anywhere near the coast by road. Most of the coast is lined with a broad and virtually impenetrable band of coastal wetlands that are far better suited for the abundant fish and wildlife of the region than for human habitation. In most coastal states, especially Florida and California, where proposals for outer Continental Shelf (OCS) development have met with intense opposition, the situation is completely different. Most of the population lives on or near the coast, and virtually all of the coast is readily accessible by road. In California, Florida, and much of the rest of the coastal United States, the coast is seen as a valuable public resource, a thing of beauty, and a source of popular recreation—in part because the coast is actually seen by so many people, so often.

Because of the coastal wetlands in Louisiana, not only is the coast rarely seen, but local residents' descriptions of their state's coastal regions are more likely to involve alligators and mosquitoes than spectacular scenery.

The second key feature of Louisiana's coastal topography is the presence of an extensive estuary system. Although it is difficult to reach the Louisiana coast from land, it is easy to do so from the water. Louisiana is characterized by numerous waterways that intersect the highway network farther inland, and that provide coastal access for marine interests. To those familiar with other coastlines, the concept of harbor is probably a common one. A harbor is generally thought of as an enclosed, protected water body with orderly docks, designed to maximize the number of boats that can be stored in the limited protected waters. Outside of the New Orleans metropolitan area, the concept is little used in coastal Louisiana, where people simply live among their boats. There is no shortage of places to put a boat in coastal Louisiana. The natural levees of the many bayous that traverse the area provide innumerable places both to dock boats and to live. People do both. The geomorphology of the Louisiana coast even allows oil field and fishing vessels to use many of the same facilities. There have been a few cases, such as Morgan City in the boom years of the 1970s and early 1980s, in which the demand for dock space became so intense that the shrimpers were crowded out. Generally, however, fishing and oil field boats have in the past, and do today, coexist peacefully. Along the coast of California, by contrast, most of the available harbors are small, and dock space is quite limited. A 250-ft offshore supply vessel represents a considerable threat to local fishing interests.

The third key feature is that Louisiana is a region of extremely low relief and generally low energy. When Louisiana residents do reach the shoreline of their state, they see flat water and flat land, with comparatively low-energy beaches. They see water that is seldom clear because of the discharge of silt by the Mississippi and Atchafalaya Rivers. In areas of the country that are farther removed from the "muddy" Mississippi, including the areas to the east in Florida, the waters are generally clearer. In northern California, residents frequently observe the power of the sea where it meets the coast, and they often comment on it with what they call "awe." In contrast, the Gulf of Mexico off of Louisiana is generally a more sedate body of water, and even when the Gulf does inspire awe, such as during tropical storms and hurricanes, its displays have

rarely been seen by people who are alive today. As noted previously, only a small percentage of the population lives on the Gulf, and a combination of modern weather forecasting and centuries of experience have led to routine evacuations of coastal and low-lying areas as tropical storms approach.

The distinctiveness of the Louisiana environment, however, also extends offshore and to what is literally below the surface—that is, the marine topography. Most people in Louisiana do not see offshore oil as an issue, and many are particularly bewildered by the fact that fishing interests in other locations have expressed so much concern over offshore developments, given that marine use conflicts have been so notably absent in the Gulf. Many fishing boats, both commercial and recreational, will actually tie up to the oil rigs to fish, and in New Orleans the Aquarium of the Americas includes a large display, complete with the title "From Rigs to Riches," pointing out the advantages of oil platforms as a form of habitat for many of the fish species of the Gulf. Given the Louisiana experience, why would there be so much opposition to oil structures from fishing interests in other regions? The reasons are not so obscure as they first appear. In addition to the historical factors previously noted, there are two key aspects of the marine environment that have helped to limit the potential for conflicts between OCS development and other marine activities, particularly fishing, off the coast of Louisiana. One has to do with the gradual slope of bottoms along the central Gulf of Mexico, and the other has to do with the presence of silt bottoms that limit the number of obstacles likely to be encountered.

Off the coast of Louisiana, the gradual slope of the Continental Shelf presents very different conditions from those found in areas where offshore oil development has been most contentious. In most areas of the country, particularly along the Pacific Ocean, the Continental Shelf drops off much more dramatically into the ocean basin, meaning that it is also much narrower than the shelf that exists in the Gulf.

As a result of this difference in sea-bottom slopes, the available area of the Louisiana Continental Shelf is far larger than is the case in California, reducing significantly the probability of use conflicts. The gradual slope also reduces the number of problems that are likely to be created by any given obstacle. Even if a fishing boat needs to make a quarter-mile detour around oil operations, for example, there will be little significant impact on the boat's ability to keep its nets in contact with the sea floor. In contrast,

in areas such as the Pacific Ocean off the California coast, the steeper slope makes the Louisiana experience largely irrelevant in two ways. First, the actual area available for use is smaller, meaning that even an apparently small loss of area for OCS activities can have a proportionately major impact on the area available for fishing in a given region. Second, given that bottom-dragging operations need to work along a contour line, following sea bottoms at a given depth, the presence of an oil platform can mean that fishing boats would need to make a detour that would correspond to a difference of several hundred feet in water depth, effectively precluding the option of “fishing around” many such structures.

An additional feature of the marine environment of the Gulf of Mexico involves the presence of silt bottoms. Although the heavy silt discharges by the Mississippi and Atchafalaya Rivers mean that the water is seldom clear, reducing many concerns over esthetic impacts, the nature of the bottom also means that fishing operations encounter few obstacles such as rock outcroppings that would lead to the loss of nets and other gear. In regions such as California, the frequent presence of rocky outcroppings can severely limit the ability of fishing boats to change their trawl runs.

One consequence of the difference in bottom types is that, to reiterate a point that often comes as a surprise to the opponents of oil development in other regions, oil rigs actually can provide a significant advantage for fishing operations on silt bottoms of the Gulf of Mexico. Certain types of commercially important fish can only survive in the kinds of habitat known collectively as hard substrate—rocky bottoms, reefs, rock outcroppings, and the like. In the central Gulf of Mexico regions where oil development activities have been the most intense, natural outcroppings of this sort are so rare along the predominantly silty bottoms that oil-related structures now comprise approximately 25% of all hard substrate. In effect, the oil rigs serve as artificial reefs, concentrating and probably increasing the fish populations, which is precisely why it is quite common to see fishing boats literally tied up to oil rigs in search of fish.

In summary, in Louisiana the coastline is inaccessible to most land-based populations, and accordingly low in social salience, while offering enough potential harbor space to meet the needs of offshore oil development as well as potentially competing uses such as fishing operations. The offshore seafloors, meanwhile, tend to have such gradual slopes as to offer vast areas of virtually level bottoms, relatively

free of obstacles but also so devoid of natural reefs that the oil rigs provide a valuable service for fishing operations. As was the case for the historical factors, most of these characteristics are almost precisely reversed for the coastal regions of much of the nation, particularly the northern California coast; the very considerations that have contributed to the ready acceptance of offshore oil in Louisiana tend to exert the opposite effect in California.

Social factors provide a third broad category of reasons why the attitudes toward offshore development can differ so widely, even within the context of a single country. As a useful simplification, four social factors appear to have encouraged the easy acceptance of offshore oil in Louisiana relative to what might be expected in other areas of the United States: the average educational levels, the patterns of social contacts, the importance of prior extractive industries, and the potential for overadaptation that characterized the coastal regions of Louisiana as offshore activities were first being developed.

Given that support for environmental protection is quite strong in the United States, as it is in most countries of the world today, very few sociodemographic characteristics show strong correlations with environmental concerns. For example, studies show that blacks are as supportive of strong environmental controls as whites, and poor people tend to be as supportive as wealthier people. One consistent exception, however, has to do with educational levels, with better educated persons generally expressing higher levels of environmental concern. Thus, it may be significant that, particularly as offshore development was picking up speed in the 1930s and 1940s, coastal Louisiana had some of the lowest educational levels in the country. In St. Mary parish, the scene of initial offshore activity, only 47.2% of the adult population had 5 years of education in 1940, and only 12.2% had graduated from high school. Other rural areas of southern Louisiana had similarly low educational levels. In comparison, more than 78% of adults in the United States had a high school education or more by the 1990 Census.

Second, the other industries that most characterized coastal Louisiana at the time of initial offshore development were ones that involved extractive uses of the coast. Like oil development, they involved the extraction of raw materials from nature. Local residents obtained products from the Atchafalaya Basin (cypress lumber, fish, crawfish, water fowl, and moss for furniture stuffing) and from the coastal marsh (furs, shrimp, and oysters). The export of such raw materials had provided the mainstay of the

economy in coastal Louisiana for almost a century prior to offshore development. In areas where extractive activities are dominant, as in Louisiana in the 1940s and 1950s, there tends to be less resistance to additional extractive activity. In most coastal regions of the United States today, and certainly in those regions of California where proposals for offshore oil have been met with the stiffest resistance, the economy has come to be far more dependent on the amenity values of the coast than on its extractive values. The proportion of the population depending on the extraction of coastal resources is small and shrinking, whereas not just tourism but also other amenity-related forms of development, including many high-value economic activities, depend on high levels of environmental amenities. These growing economic interests place high importance on maintaining relatively pristine environmental conditions. Thus, in California's coastal areas, and increasingly in other areas as well, the likelihood of finding support from extractive workers can be expected to continue to decline.

Third, social interaction patterns can exert powerful influences on individuals' attitudes. One particularly powerful, if often overlooked, pattern involves what we have called a social multiplier effect: Even if a given individual does not work in the offshore oil industry, that person's attitudes may be affected by whether his or her friends and relatives do. Given the historical, biophysical, and other social factors summarized previously, the "average" resident of coastal Louisiana in the 1940s would be expected to have known many friends and neighbors who were employed in the oil industry. By the 1980s or 1990s, it was virtually impossible to live in southern Louisiana and not know someone who was employed in an oil-related enterprise. In addition, as reported in the 2000 Census, Louisiana residents were more likely to live in the state in which they were born than were residents of any other state. In contrast, in most of the coastal regions of the United States today, the situation is the opposite. Not only do these regions have few residents who are involved in oil or gas extraction but also they have high numbers of people who have moved to the coastal regions precisely because of their desire for high environmental quality. This tendency is especially likely in areas such as northern California, where there is no current base of employment directly dependent on local oil and gas development.

Finally, an additional way to understand the uniqueness of Louisiana's compatibility with offshore development is to examine the degree of

adaptation that has taken place. As the various components of the human environment become adapted to a given form of development activity, there is a tendency for new skills, knowledge, tools, networks, and other resources to be built up around that activity. Three potentially problematic results, however, can also occur. First, any type of development narrows a region's options because time, resources, and human capital are devoted to a particular developmental scenario, sometimes limiting the options for alternative scenarios. We call this process developmental channelization, although others have used terms such as path dependency to denote similar patterns. Second, as part of this process, adaptations to earlier forms of development may be lost (sometimes quickly and sometimes across generations). Third, nearby communities can run the risk of overadaptation to the new development activity. Particularly if that new activity is not a sustainable one, then when it ceases or declines, communities or regions may be left in the position of being less able to survive in their environment than they would have been before the new development occurred. Thus, to a large extent coastal Louisiana

TABLE I

Factors That Led to Positive Evaluations of Offshore Oil in Louisiana^a

Historical factors	
	Early historical era—"age of exuberance"
	Temporal priority of development
	Incremental onset and evolution of industry
	Local origins and development of technology
Biophysical factors	
	Coastal topography
	Broad coastal marshes that preclude coastal highways and population concentrations
	Estuary systems that include many potential harbors
	Low relief and energy levels
	Marine topography
	Broad, gradual slopes that increase area and decrease significance of conflicts with platforms
	Silt bottoms with few outcroppings and obstacles
Social factors	
	Low educational levels
	Extractive orientation toward biophysical environment
	Favorable patterns of contact with oil industry personnel
	Extensive prior adaptation to oil development

^a Adapted from Freudenburg and Gramling (1994).

has become not just adapted to but also dependent on its offshore petroleum activities, and it is difficult to be too critical in evaluating an activity from which one has benefited, particularly if the activity is one on which one remains dependent for any hopes of comparable future prosperity. This, coupled with the industry's generally good record in terms of offshore environmental accidents, has tended to bolster the perception of offshore activity in the Gulf as involving relatively low levels of risk. Table I summarizes these factors. As a simplified but generally helpful rule of thumb, where more of these factors are present, attitudes toward offshore development can be expected to be more positive.

SEE ALSO THE FOLLOWING ARTICLES

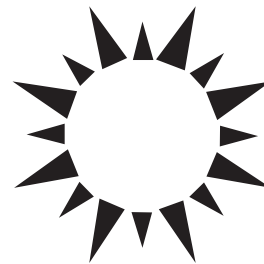
Crude Oil Releases to the Environment: Natural Fate and Remediation Options • Oil and Natural Gas Drilling • Oil and Natural Gas: Offshore Operations • Oil-Led Development: Social, Political, and Economic Consequences • Public Reaction to Electricity Transmission Lines • Public Reaction to Energy, Overview • Public Reaction to Nuclear Power Siting and Disposal • Public Reaction to Renewable Energy Sources and Systems • Wetlands: Impacts of Energy Development in the Mississippi Delta

Further Reading

Ahmad Khan, S. (1994). "Nigeria: The Political Economy of Oil." Oxford Univ. Press, Oxford, UK.

- Andersen, S. S. (1993). "The Struggle over North Sea Oil and Gas: Government Strategies in Denmark, Britain, and Norway." Scandinavian Univ. Press/Oxford Univ. Press, New York.
- Baldwin, P. L., and Baldwin, M. F. (1975). "Onshore Planning for Offshore Oil: Lessons from Scotland." Conservation Foundation, Washington, DC.
- Barnes, P. (1995). "Indonesia: The Political Economy of Energy." Oxford Institute for Energy Resources, Oxford, UK.
- Freudenburg, W. R., and Gramling, R. (1994). "Oil in Troubled Waters: Perceptions, Politics, and the Battle over Offshore Drilling." State Univ. of New York Press, New York.
- Gramling, R. (1996). "Oil on the Edge: Offshore Development, Conflict, Gridlock." State Univ. of New York Press, New York.
- Gramling, R., and Freudenburg, W. R. (1996). Crude, coppertone, and the coast: Developmental channelization and the constraint of alternative development opportunities. *Soc. Nat. Resour.* 9, 483–506.
- Jones, R. E., and Dunlap, R. E. (1992). The social bases of environmental concern: Have they changed over time? *Rural Sociol.* 57(1), 28–47.
- Molotch, H., Freudenburg, W. R., and Paulsen, K. (2000). History repeats itself, but how? City character, urban tradition, and the accomplishment of place. *Am. Sociol. Rev.* 65(6), 791–823.
- Okome, O. (ed.). (1999). "Before I Am Hanged: Ken Saro-wiwa—Literature, Politics, and Dissent." Africa World Press, Trenton, NJ.
- Ross, M. L. (2001). Does oil hurt democracy. *World Politics* 53, 325–361.
- Seda, F. S. S. E. (2001). "Petroleum paradox: Natural resources and development in Indonesia, 1967–1997." Unpublished PhD dissertation, University of Wisconsin, Madison.
- Seyfrit, C. L. (1988). A need for post-impact and policy studies: The case of the Shetland Experience. *Sociol. Inquiry* 58, 206–215.
- Smith, D. A. (1994). Uneven development and the environment: Toward a world-system perspective. *Humboldt J. Soc. Relations* 20(1), 151–173.
- Storey, K. J., and Shrimpton, M. (1987). "Planning for Large-Scale Construction Projects: A Socio-economic Guide for Communities, Industry and Government." Environmental Studies Research Funds, Ottawa, Ontario, Canada.

Public Reaction to Renewable Energy Sources and Systems



TIMOTHY C. COBURN

Abilene Christian University
Abilene, Texas, United States

BARBARA C. FARHAR

University of Colorado
Boulder, Colorado, United States

-
1. Definition of Renewable Energy
 2. Political, Philosophical, and Economic Perspectives
 3. The Evolving Consumerism toward Renewable Energy
 4. Government Support for Renewable Energy Initiatives
 5. Trends in the Demand for Renewable Energy and Renewable Energy Systems
 6. Consumer Attitudes toward Renewable Energy and Renewable Energy Systems
 7. Stated Willingness to Pay for Electricity Derived from Renewable Energy
 8. Gauging Actual Participation of Consumers in Renewable Energy Programs
 9. Public Response to Renewables in the Transportation Energy Sector
 10. Factors That Motivate Public Response to Renewables
 11. Modeling Public Response to Renewables
 12. Conclusion

Glossary

biomass Organic materials such as woodchips, twigs and stems, grasses, weeds, aquatic plants and other quick-growing plants, and animal wastes that can be used to produce fuels or electricity through a process of fermentation, distillation, or incineration.

cogeneration The simultaneous production of several types of energy; a process that converts a fuel into both thermal and electrical energy.

ethanol Nontoxic alcohol produced from sugar, starch, or other carbohydrates through the process of fermentation; sometimes referred to as grain alcohol or ethyl alcohol; used as a fuel or fuel additive to improve oxygenation.

fuel cell A device that converts the energy of a fuel directly into electricity and heat without combustion but through the process of oxidation; originally developed

for the space program and now largely intended for use in automobiles.

geothermal energy Energy in the form of natural heat originating from within the earth and contained in rocks, water, brines, or steam.

green power Electricity produced from renewable sources or in such a way that the process results in little or no adverse impact on the environment.

green pricing Strategies and formulas used by utility companies to recoup the costs of producing electricity from renewable sources, often by charging customers premiums or incremental amounts on their monthly utility bills; customer participation in green pricing programs is voluntary.

grid-tied photovoltaics (GPV) Photovoltaic systems designed to operate in parallel with and interconnected to the electric utility grid.

hydroelectric power or hydropower Energy produced by the force or pressure of falling water.

landfill gas A gas consisting mostly of methane that is generated as a by-product through the natural decomposition of organic material at landfill disposal sites.

methanol A toxic alcohol derived largely from natural gas or coal; sometimes referred to as the simplest alcohol, and other names include methyl alcohol, carbinol, and wood alcohol. It is used as an antifreeze, solvent, or fuel/fuel additive.

nuclear fission The process of splitting the nucleus of an atom into two parts, either spontaneously or as the result of an impact from another particle, resulting in the release of energy.

ocean energy Energy produced by the force or pressure of wave action.

photovoltaics (PV) Passive devices constructed from semiconducting materials that absorb sunlight and convert it into heat or electricity.

renewable energy Heat, electricity, or power produced from renewables or renewable sources, such as

wind, the sun, subsurface steam, and water/wave action.

renewables Naturally occurring resources or materials used in energy, fuel, or electricity production that can be regenerated or replenished over time.

solar energy Energy from the sun that is converted into heat or electricity through passive collector devices.

sustainable In the context of energy resources, consumption in a manner that does not result in depletion and/or use of natural assets that are renewable; in the context of the environment, usage or treatment of Earth and its natural and physical assets in a manner that does not deplete, damage, or destroy.

wind energy Heat or electricity produced by the force or pressure of wind or air currents.

Whereas most energy streams are derived from sources that are exhaustible (e.g., fossil fuels), renewable energy is produced from sources that can be replenished over time. Faced with declining availability of some nonrenewable resources and with growing concerns about their environmental and public health impacts, society has become increasingly more attracted to renewable energy that exists in greater abundance and has minimal adverse physical effects. In response to a measurable shift in public opinion over the past two decades, many initiatives have been undertaken to expand the use of renewable energy sources and systems. The results of numerous polls and surveys indicate that citizens throughout the industrialized world now desire greater reliance on renewable and are willing to invest in their development.

1. DEFINITION OF RENEWABLE ENERGY

Renewable energy means different things to different people. Although there is little argument as to what energy is, even in its myriad forms, the term renewable energy conjures up a more diverse assortment of images.

In the simplest terms, energy is the capacity of a physical system to do work, or alternatively, to produce heat. Energy is a characteristic of the physics of a system. Renewable, on the other hand, pertains to the origin of the energy; thus, ideas and opinions about renewable energy are often confounded with information (or misinformation) about the source.

Renewable energy is energy that is derived from a supply that is constantly and naturally replenished over a relatively short time. Hence, any discussion of renewable energy is ultimately reduced to a discus-

sion of renewable resources as they are derived from geophysical processes: sunlight, wind, falling water, sustainable biomass, wave motion, tides, and geothermics. Some entities, such as the Texas Renewable Energy Industries Association, describe all renewable energy as being directly or indirectly derived from the sun or natural movements and mechanisms of the environment. Others—private, commercial, governmental, and international—have their own definitions. One of the more visible characterizations has been provided by the U.S. Department of Energy (DOE), which segregates renewable energy into six components: bioenergy derived from biomass resources; geothermal energy derived from the earth's own heat-producing processes; hydrogen, hydropower, or hydroelectric power; ocean energy; solar energy; and wind energy.

2. POLITICAL, PHILOSOPHICAL, AND ECONOMIC PERSPECTIVES

Given the absence of a uniform definition, it is not surprising that the larger energy community is somewhat unclear about the role renewable energy plays in the broad arena of energy policy and economics, and that markets for renewable energy and related systems have yet to fully mature. In addition, because public perception is so often impacted by media coverage, or the lack of it, uncertainty and misunderstanding about renewable energy among the general public might be a perfectly legitimate response.

Renewable energy evokes many different metaphors and perceptions. Some people think of renewable energy as a nondepletable, environmentally friendly alternative to conventional fossil fuels, whereas others consider it to be an expensive, high-tech, and possibly unreliable proposition. Some feel threatened by renewable energy and view it as detrimental to current energy markets and jobs, but others (e.g., residents of California) believe it represents a sound investment for long-term economic and energy security.

For many years, renewable energy advocates have been unjustifiably associated with the more radical component of the environmental movement. Although this impression has moderated, the political and economic struggles over energy supply, demand, efficiency, and sustainability continue even today in countries throughout the world, and the role to be played by renewable energy, along with the concept of energy efficiency, is at the very heart of many of them.

One has to look no further than the ongoing conflict in the United States over drilling for oil in the Alaska National Wildlife Reserve (ANWR) to understand what is at stake with regard to jobs, markets, and the domestic economy. Hence, within the greater energy complex, the views and claims of renewable energy and energy efficiency advocates have sometimes been mistrusted—even totally dismissed—because, on the surface, they have seemed counter rather than complementary to conventional energy policy and wisdom. However, from a retrospective viewpoint, there has been a not-so-subtle shift in public and corporate attitudes toward renewable energy, particularly in the past decade. In the historical arena in which plentiful, inexpensive fossil energy was taken for granted, renewable energy was not even on the map. However, in the more contemporary worldview that is tempered by international conflict, global warming, economic uncertainty, environmental and public health concerns, and rolling electricity black-outs, renewable energy has a more prominent place at the table, even among those who are steeped in the fossil fuel tradition. As is the case for many things, it is a matter of perspective. It is the changing perspective regarding renewables within the energy community that holds such great interest from both the social and economic standpoints.

3. THE EVOLVING CONSUMERISM TOWARD RENEWABLE ENERGY

The posture of consumers is unquestionably changing. Although some politicians and corporate executives would still have the public believe that renewable energy is a bad investment, the more pragmatic and economically savvy consumers of the early 21st century are not buying the argument. It seems they are not quite as ill informed as they were once thought to be. First detected in the late 1980s, a distinctively proactive stance toward energy derived from renewable resources is now readily apparent in the results of almost all public opinion energy research. This change, which has been described as “the quiet revolution in renewable energy,” has occurred at the grassroots consumer level with little fanfare and surprising determination, catching the corporate energy establishment somewhat off guard. Despite earlier technological missteps, consumers now see reliance on alternative energy sources and systems as fundamental to maintaining the social norms and economic well-being of society. Consumers seem poised to rapidly adapt their lifestyles to a wide range

of alternative energy technologies, from wind turbines to fuel cell vehicles and rooftop solar electric panels.

Some of the attitudinal and behavioral change is undoubtedly due to initiatives by consumers to seek more balance in their lives, a greater degree of economic security, and more independence and control over their own futures. In addition, the post-baby-boom generation has simply been exposed to more technological innovation, and individuals in this age group have come to understand that technological expansion is a major economic force throughout the world.

4. GOVERNMENT SUPPORT FOR RENEWABLE ENERGY INITIATIVES

The change is also partly attributable to substantial government investment in research and development activities that serve not only as the catalysts for new ideas and equipment but also as incubators for fledgling industries that create jobs. Through government mandates, rebates, tax credits, and other incentives, use of renewable energy and renewable energy systems is proliferating at both the local and the national levels. In the United States, all federal agency vehicle fleets must now contain specified percentages of alternative fuel vehicles, and energy consumption at federal facilities and installations must include specified percentages from renewable sources. As an example of the results of such a proactive government stance, late in 2002 Dyess Air Force Base in Abilene, Texas, became the first U.S. facility to be fully dedicated to wind power. Programs, situations, circumstances, and events of this nature, being publicized through the media, definitely have a positive, although perhaps unquantifiable, impact on consumer awareness and behavior.

Throughout the United States and in other countries as well, government agencies have engaged in a fairly robust educational blitz aimed at promoting the development and use of renewable resources. Such educational initiatives begin at the elementary school level in an effort to establish an attitude toward renewable energy that gives rise to lifelong patronage and support. Numerous advocacy groups, such as the Colorado Renewable Energy Society, the American Council for Renewable Energy, the Clean Energy Funds Network, and Citizens for Renewable Energy in Canada, have taken up the charge and play a major role in the overall educational process. Increasingly more public and grassroots community events, such as the 2003 Renewable Energy Roundup

and Green Living Fair in Fredericksburg, Texas, and the 2004 Renewable Energy and Sustainable Living Fair in central Wisconsin, are being organized and promoted, although the total number of individuals who participate in these activities and their reasons for doing so have not been researched and are unknown.

In the United States, the federal government also provides grants directly to individual citizens and small businesses to encourage the use and development of renewable energy. The U.S. Department of Agriculture sponsors a grant program to assist farmers, ranchers, and rural small businesses in developing renewable energy systems and energy efficiency improvements. The DOE has also announced grant awards to eight Native American tribes to develop renewable energy projects on their lands.

In addition, state and regional governments have been prime marketers of information about renewables. Many, if not all, U.S. state energy offices actively promote the use of renewable energy through Internet sites and other media. The events in California during the past few years have brought renewable energy to the forefront and have impacted public opinion about energy (at least in that area of the United States) in ways that might not otherwise have been possible.

5. TRENDS IN THE DEMAND FOR RENEWABLE ENERGY AND RENEWABLE ENERGY SYSTEMS

With all this activity and an increasingly pro-renewables consumer stance, one might expect actual use of renewable energy to be skyrocketing. In some

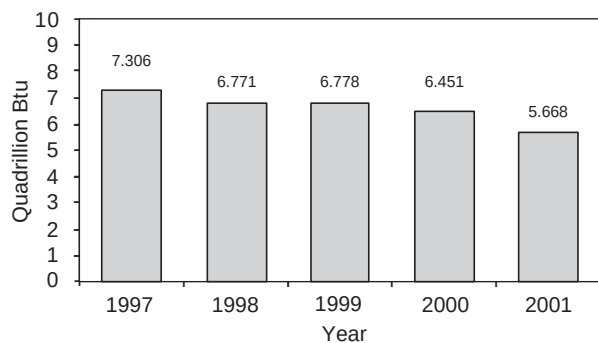


FIGURE 1 Total U.S. consumption of renewable energy, all sources, 1997–2001. Data for 2001 were preliminary at the time of reporting. Data from the U.S. Energy Information Administration (2002b).

ways, this is true, but in others it is not. The U.S. Energy Information Administration (EIA) reports in its “Renewable Energy Annual” that in 2001, total renewable energy consumption (solar energy, biomass, geothermal energy, conventional hydroelectric, and wind energy) declined by more than 12%, resulting in a decrease in the share of total U.S. energy consumption to 6%, the lowest level in 12 years (Fig. 1). These percentages are somewhat misleading because much of the change is due to a sharp reduction in the use of hydropower resulting from environmental and political pressures to eliminate dams in the Pacific Northwest. Shipments of solar thermal collectors, on the other hand, surged from 8.4 million square feet in 2000 to 11.2 million square feet in 2001 (Fig. 2), whereas total renewable electric generating capacity increased by a modest 1803 MW to 96,741 MW in 2001 (Fig. 3) due largely to increased wind power generation. The latest data available on total shipments of geothermal heat pumps indicate that they declined from 41,679 in 1999 to 35,581 in 2000 (Fig. 4). A new DOE initiative pertaining to geothermal heat pumps, planned to begin in 2004, may reverse this trend.

A major reason why renewable energy consumption has not exhibited larger increases is that utility

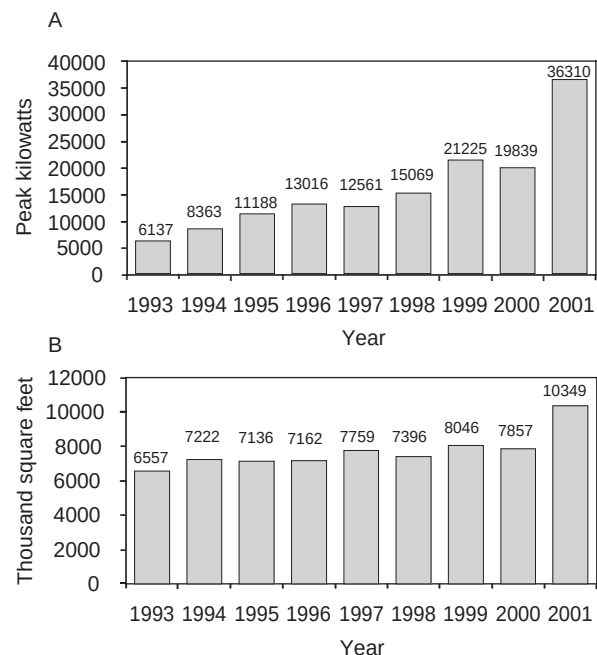


FIGURE 2 Annual total domestic shipments (U.S. states and territories, excluding shipments for space/satellite applications) of photovoltaic cells and modules (A) and solar thermal collectors (B), 1993–2001. Total shipments are net of exports. Data from the U.S. Energy Information Administration (2002b).

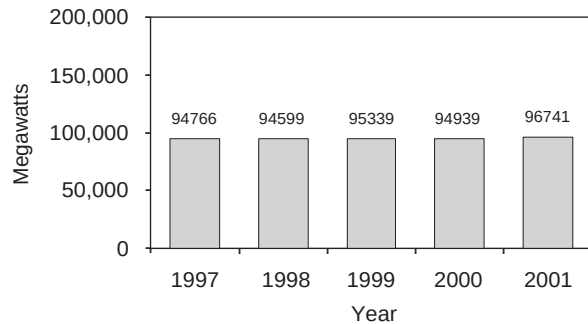


FIGURE 3 United States. electric net summer capacity attributable to renewables, all sources, 1997–2001. Data for 2001 were preliminary at the time of reporting. Data from the U.S. Energy Information Administration (2002b).

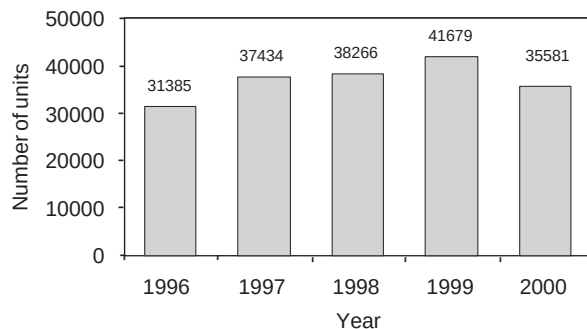


FIGURE 4 Annual total shipments of geothermal heat pumps, 1996–2000. Data from the U.S. Energy Information Administration (2002b).

companies have been slow to incorporate renewables into their resource base, citing economic factors associated with developing infrastructure and reliable resource supplies. Another reason is that utility regulators have been slow to adopt pro-renewable energy policies, and local government agencies have been inconsistent in their pursuit of economic incentives that promote market development. Although it is true that sufficient infrastructure may not yet exist to support continuous growth in the renewable energy market, this is really a proverbial “chicken-and-egg” problem. Markets cannot develop unless investments are made, and investments cannot be made unless the market is thought to be of sufficient size to justify them.

The utilities, however, being driven by their own consumer research, are finally getting the message. Utility customers are telling their electricity suppliers that they want renewables in the resource base. During the past few years, approximately 85 utilities in the United States have begun offering electricity generated from wind and solar power to their

residential and commercial customers. The DOE provides a variety of information in both written and electronic form about successful ventures of local utilities into renewable energy. As a result, two phenomena are occurring: Energy supply is becoming more diverse and less centralized, and the demand for microturbines, rooftop-mounted solar panels for water heating and electricity generation, and other decentralized power sources is intensifying, in both the residential and the commercial markets.

6. CONSUMER ATTITUDES TOWARD RENEWABLE ENERGY AND RENEWABLE ENERGY SYSTEMS

Public response to renewable energy sources and systems is commonly measured through market and opinion surveys. Many such surveys pertaining to renewable energy have been conducted during the past 20 or more years by various organizations with different agendas. The quality of such surveys is highly variable due to inconsistency in survey format, question phrasing, and selection of respondents. Attitudinal responses are also affected by respondents’ knowledge about specific energy sources since respondent knowledge is often uneven. In addition, almost all the results are descriptive in nature and provide little in the way of inferences with regard to the real populations of interest. Furthermore, consumer attitudes about renewable energy can span a wide variety of topics, so the research results are not always directly comparable. Table I provides comparative information about the operational and statistical aspects of a cross section of market and opinion research efforts at the local, regional, and national levels that pertain to renewable energy. Data obtained in objective, national surveys are thought to be the most reliable and unbiased.

Despite the acknowledged difficulties with some surveys, a trend has emerged with regard to consumer attitudes toward renewable energy. Some of the most telling results have been obtained from utility market surveys, many of which have been conducted by and for individual utility companies. Two 1999 reports summarize some of the findings. The first report, prepared at the National Renewable Energy Laboratory (NREL), contains an aggregate summary of results obtained in 14 different “best practices” quantitative surveys encompassing 12 utility service territories in five western/southwestern

TABLE I

Comparative Information on the Operational and Statistical Aspects of a Cross Section of Market and Opinion Research Efforts (Regional, Local, and National) Pertaining to Renewable Energy, 1995–2002

Year	Survey organization	Sponsor	Thematic content ^a	Population coverage	Subjects ^b	Survey mode	Margin of error (%) ^c
1995	Research/Strategy/Management, Lanham, MD	Sustainable Energy Coalition, Takoma Park, MD	A	Nationwide	1000	Telephone	3.10
1997	Action Marketing Research, Minneapolis, MN	Environmental Law & Policy Center, Chicago, IL	A	Minnesota	900 ^d	Telephone	Unknown
1997	Mellman Group, Washington, DC		A	Nationwide	800	Telephone	3.50
1998	The Eppstein Group, Ft. Worth, TX	Houston Lighting & Power, Houston, TX	A	Houston	> 200	Deliberative poll	N/A
1998	Center for Deliberative Polling, University of Texas, Austin	TU Electric, Texas	A, B	TU Electric customers	Unknown	Deliberative poll	N/A
1998	Center for Deliberative Polling, University of Texas, Austin	Union Pacific Resources, Ft. Worth, TX	A	Texas	1207	Unknown	<3.0
1998	The Eppstein Group, Ft. Worth, TX	Southwest Public Service (SPS), Amarillo, TX	A, B	SPS customers	Unknown	Deliberative poll	N/A
1998	Research/Strategy/Management, Rockville, MD	Sustainable Energy Coalition, Takoma Park, MD	B	Nationwide	1003	Unknown	3.10
1998	The Texas Poll, Scripps Howard and Office of Survey Research, University of Texas, Austin	Sierra Club, Lone Star Chapter, Austin, TX	B	Texas ^e	1004	Telephone	3.00
2000	International Consumer Research, Media, PA	Deloitte & Touch, New York, NY	A	Nationwide	600 ^f	Unknown	Unknown
2001	Mellman Group, Washington, DC		B	Nationwide	1000	Telephone	3.10
2001	The Gallup Organization, Lincoln, NE		B	Nationwide	512 ^g	Telephone	5.00
2001	Iowa State University, Ames, IA		B	Iowa farm producers	2060 ^b	Unknown	Unknown
2002	Opinion Dynamics Corporation, Cambridge, MA	MA Technology Collaborative, Boston, MA	A, B	Massachusetts ⁱ	650	Telephone	3.84
2002	ORC/Macro Intl, Calverston, MD	NRG Systems, Hinesburg, VT	B	Vermont	402	Telephone	Unknown

^a A, willingness to pay for green pricing; B, energy policy issues or emphasis of utilities on renewables.

^b Unless specified otherwise, simple random sampling of registered voters is presumed.

^c Usually stated as an overall rate, with subgroup rates being higher. Based on a statistical confidence level of 95%.

^d Equal allocation among some counties and cities.

^e Adults over the age of 18.

^f Consumers (undefined).

^g Adults.

^h Based on multiplying reported response rate, 61%, times the original sample size (3376).

ⁱ Massachusetts residents.

states (results from an Electric Power Research Institute study are also included). Drawing on the approach and research used at NREL, the second report, prepared at the Bonneville Power Administration (BPA), contains a similarly constructed summary of 17 surveys or focus group studies conducted in the Pacific Northwest.

Included in the NREL report are the results for representative questions abstracted from the 14 surveys. One such question asked the customers to indicate the extent to which they favor various energy resources. The percentages of respondents indicating that they “somewhat favor” or “strongly favor” solar energy, wind, natural gas, geothermal energy, nuclear fission, and coal were 93, 91, 83, 71, 31, and 24%, respectively, suggesting a higher degree of favorability toward renewables than more traditional energy sources. Another question from this group of surveys asked customers to indicate which energy sources their utilities should use to generate power. Of the residential customers that responded, 88% said their utilities should use renewables (solar, wind, geothermal, or landfill gas) and 85% said their utilities should use hydropower, whereas only 55, 39, and 14%, respectively, said their utilities should use natural gas, nuclear fission, and coal. In a third question, customers were asked to indicate the source from which they preferred their electricity to be generated. Of those responding, 41% said they would choose wind and solar energy as their electricity source, 35% said they would choose natural gas, 9% said they would choose nuclear fission, and only 5% said they would choose coal.

Similar results are reported in the BPA document. In a question included in the 1997 Environmental Issues and Tracking Study commissioned by Portland General Electric and summarized by BPA, customers were asked to indicate the single additional energy source they would like to see their electric company develop during the next 5–10 years. Sixty-nine percent chose solar, wind, geothermal, or hydroelectric, 11% chose cogeneration using waste heat, 9% selected natural gas-powered turbines, 2% selected fuel cells, 2% chose coal, and 7% indicated some other source or did not know. The BPA report also summarized the results from a 1997 survey commissioned by Conservation and Renewable Energy Systems, an energy and conservation consortium of public utilities in Washington State. In that study, 88 and 89% of residential and commercial customers, respectively, indicated that utilities should include renewables as one of their energy sources. Even when residential customers were told

that “renewables may cost more today but might make sense in the long run,” there was no change in the percentages favoring them.

These comprehensive assessments reveal a repeated and consistent pattern that strong majorities of utility customers in various regions of the United States prefer renewables to other available energy sources and favor utility use of renewable resources to produce electricity. This pattern corroborates findings in polls conducted at the national level during a period of 23 years that Farhar summarized and described in a 1996 publication for the Renewable Energy Policy Project (REPP) in Washington, DC.

Some of the most recent evidence about consumer attitudes toward renewables was obtained in a 2002 survey conducted by the Wells Fargo Public Opinion Research Program (Wells Fargo PORP) at the University of Colorado, Denver. In this case, 95% of respondents disagreed with the statement, “It really makes no difference to me how my utility generates electricity.” Of these same respondents, 82% indicated that utilities should focus on renewable resources as opposed to fossil resources to generate additional electric power; and when asked which source they would most like to see developed, 82% chose a renewable one (wind, solar, or hydropower). Furthermore, 84% indicated that the state’s utilities should use “somewhat more” or “a lot more” renewable energy than they currently are using, 62% indicated that the state legislature should require utilities to use more renewable energy, and 78% said they would support a legislative mandate for utilities to produce 10% of their electricity from renewables within the next 10 years. Finally, strong majorities of respondents in this study rejected the notions that (i) it is too soon for utilities to invest in renewable energy, (ii) renewable energy is not worth developing because there is not enough of it, and (iii) wind farms are ugly.

The opinions and attitudes of European consumers toward renewable energy appear to mimic those of citizens in the United States. In a 2002 survey conducted by the European Commission, 69% of respondents said they supported more research for renewable energy, whereas the percentages of respondents who said they supported more research for natural gas, oil, and coal were only 13, 6, and 5%, respectively. On the other hand, only 40% stated they believe renewable energy would be the least expensive source in 2050. Like so many surveys of its type, this research does not directly address consumer attitudes toward actual use of renewables in power generation. It does, however,

serve to highlight the tendency of consumers to favor renewables when all other things are equal.

Information on consumer acceptance of specific renewable energy technologies and systems is more limited. Perhaps the most recent and extensive study about a particular technology, grid-tied photovoltaics (GPV), was conducted at NREL in 2000. In this survey of homeowners in Colorado, 68% of respondents favored GPV being made more widely available, but there was clearly greater favorability toward than familiarity with the technology. In the same study, participants were asked to identify which electricity source—coal, hydropower, natural gas, nuclear fission, oil, or solar energy and wind—they think of most when presented with each of several descriptive statements. Table II indicates the percentage of respondents who selected solar energy and wind with respect to each of the statements, providing information on a somewhat broader spectrum of attitudes toward renewable energy than presented in the other work.

Several surveys have been conducted in the United Kingdom concerning consumer favorability toward wind energy development. In one such survey, the University of Wales surveyed public opinions on behalf of the British Broadcasting Corporation at three different sites. At all three sites, majorities of greater than 60% of participants responded in favor

of wind development despite aesthetic concerns associated with equipment installation. Environmental superiority to other fuels was most often selected as the main attribute of wind energy.

Collectively, this evidence suggests that a sizable percentage of consumers are favorably disposed toward renewable energy, particularly from the environmental standpoint; consumers and utility customers want to see more renewable energy use; and markets for renewable energy systems are growing.

7. STATED WILLINGNESS TO PAY FOR ELECTRICITY DERIVED FROM RENEWABLE ENERGY

Although measuring consumer attitudes and perceptions about renewable energy is relatively straightforward, quantifying actual program participation or likely purchases of renewable energy and associated equipment is not. A consumer's proactive stance toward renewable energy does not always lead to direct or explicit actions that represent true consumption behavior. The reasons are numerous and complex, although some are more apparent than others. A full discussion is beyond the scope of this article.

One of the more obvious reasons is that the cost of producing electricity from renewable resources may be more than the cost of producing it from fossil fuels, and so the prospect of higher monthly utility bills sometimes dampens the enthusiasm of consumers despite their personal preferences. Consequently, a disparity exists between consumers' stated preferences for renewables or green pricing programs and their actual conduct/behavior when it comes time to participate or financially underwrite resource and infrastructure development.

Potential consumer actions are difficult to quantify directly. Consequently, stated "willingness to pay" is often used as a surrogate indicator of likely behavior in many avenues of consumer research. Although imperfect as a measure of actual behavior, willingness to pay is a notion that addresses the idea of potential consumer response to technological enhancements and it allows businesses, industries, and governing bodies to determine the probable extent to which consumers might embrace products and technologies in the marketplace. Willingness to pay, particularly in the context of green pricing, has been used extensively in energy market and opinion

TABLE II

Percentages of Colorado Respondents to a Survey about Grid-Tied Photovoltaics Who Associate Wind and Solar Energy with Several Economic and Environmental Statements

Statement	Wind and solar (%)
Best for the environment	79
Safest	70
Most abundant	43
Makes the United States most self-reliant	35
Least expensive	24
Will help the most in meeting the growing demands for energy	41
Stimulates economic development the most	24
Diversifies and extends our energy mix the most	52
Improves the quality of our air the most	80
Adds the most new high-technology jobs	40
Use of this energy source will increase the most in the 21st century	60
Best for the U.S. economy	32
Most positive overall reaction	51

Source: Farhar and Coburn (2000).

surveys during the past decade as a proxy for consumers' actual purchase of electricity generated from renewable resources.

The 1999 NREL report described previously provides an authoritative summary of stated willingness to pay for electricity from renewable resources. As previously noted, this work encompasses the results of 14 different surveys and an Electric Power Research Institute study. In all the surveys reviewed, majorities of respondents indicate they are willing to pay at least a modest incremental monthly amount for electricity generated from renewable resources. The sizes of these majorities range from 52 to 95% without prior coaching or training regarding the economic or technological issues. The percentages are even higher when such information is made available.

Based on the similarity in the nature of the survey findings, an aggregate residential customer willingness-to-pay curve can be developed representing the incremental monthly amount that consumers would be willing to pay for electricity produced by their utility companies from renewable sources (Fig. 5). As expected, the percentage of consumers who are willing to pay more decreases sharply as the amount increases. Nonetheless, based on this analysis, approximately 70% of consumers could be expected to report a willingness to pay at least \$5 more per month for electricity produced from renewable resources, and approximately 38% of consumers could be expected to report a willingness to pay at least \$10 more per month. Although willingness to

pay is not the same as actual purchase, these percentages are not trivial and they represent fairly solid market potential.

The results pertaining to stated willingness to pay from the utility-sponsored surveys summarized in the 1999 NREL report corroborate other survey results. Farhar's 1996 report on behalf of REPP indicates that 57–80% of the respondents in national polls state they are willing to pay more for "cleaner" electricity or electricity generated from "environmentally friendly" sources. The 2000 NREL study of Colorado homeowners pertaining to GPV indicates that 76% of respondents are willing to pay at least \$1 more per month for electricity generated from renewable sources. Further corroborating evidence from polls of adults in four midwestern states is reported in a 1998 study conducted by Tarnai and Moore at Washington State University, along with information contained in the 1999 BPA report.

Some results that specifically pertain to wind power program participation are presented in a 1998 report by the Energy Center of Wisconsin. In contrast to the opinion and market research surveys described in the 1999 NREL and BPA studies, this one is more experimental in nature, the research is more controlled, and the analysis is more rigorous. Of the 761 respondents who were asked hypothetically if they would be willing to pay an additional amount on their monthly utility bill to obtain wind power, 43% said yes. Of these, 61% said they would be willing to pay at most an additional \$2 per month or \$24 per year. Fifty-eight percent said they would be willing to pay at most \$4 more per month or \$48 per year, 50% said they would be willing to pay at most \$8 more per month or \$96 per year, and 18% said they would be willing to pay at most \$12 more per month or \$144 per year. These results compare closely with those noted previously (Fig. 5).

Perhaps the most recent research in the United States on willingness to pay for renewables is found in the study conducted by Wells Fargo PORP. In this study, 55% of respondents indicate they are willing to pay an additional \$10 or more on a typical monthly bill of \$55 for electricity generated from renewables, whereas 18% are willing to pay nothing more. Two-thirds (67%) of the respondents indicate that any additional cost of using renewables to generate electricity should be spread across all consumers.

Current knowledge about U.S. consumers' stated willingness to pay for renewables closely parallels research findings in other industrialized countries, although more information seems to be available for

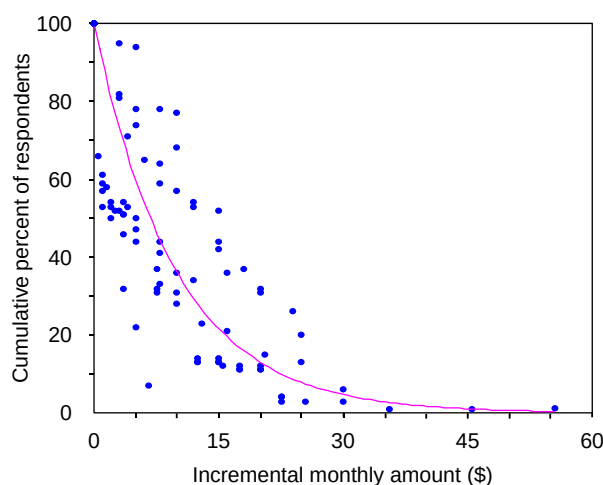


FIGURE 5 Aggregated willingness-to-pay curve for residential customers. The equation of the curve is $Y = 100e^{-104M}$, where Y is the cumulative percentage of respondents and M is dollars more per month ($R^2 = 0.76$). Source: Farhar (1999).

the United States. For example, in a small survey conducted on behalf of New Zealand's Energy Efficiency and Conservation Authority, more than half of the respondents indicated a willingness to pay 5% more for wind-generated electricity, and approximately 20% were willing to pay an additional 15%.

Irrespective of issues about comparability, the combined message from all these studies is that a high percentage of consumers nationwide, and perhaps even in other areas of the world, are indeed willing to pay more on a monthly basis for electricity produced from renewable sources. Of all possible indicators, this one signifies a fairly positive public response to renewable energy sources and systems in the general context of green pricing.

The finding suggests that markets for renewable energy and renewable energy systems do exist and that they can be expanded through strategic planning and implementation.

8. GAUGING ACTUAL PARTICIPATION OF CONSUMERS IN RENEWABLE ENERGY PROGRAMS

As suggested previously, engaging consumers in renewable energy participation has been the subject of much thought and investigation. However, because it is largely proprietary in nature, information about actual consumer participation in renewable energy programs and their purchases of renewable energy systems is difficult to obtain and assess. There are both operational and technical difficulties to overcome as well as definitional inconsistencies. For several years, the EIA has tracked shipments of solar PV cells/modules and solar thermal collection devices as a proxy for consumer demand (both private and commercial); as noted previously, these shipments were at an all-time high in 2001. Such equipment includes solar hot water devices. In fact, the EIA reports that in 2001, the residential sector replaced the industrial sector as the largest market for PV cells and modules. This evidence alone suggests, at least indirectly, that consumer demand for renewable energy systems is on the rise.

In addition, since April 2000, NREL staff have compiled and reported a "top ten" list of green pricing programs in the United States in each of the following categories: renewable energy sales, total number of customer participants, and customer participation rates. More than 300 utilities in 32 states offer green pricing programs. In December

2002, the number of patrons in the 10 largest programs with respect to customer participation ranged from 5644 to 72,732, for a total of 201,305 (approximately 42,000 of which are low-income customers in Los Angeles who do not pay a premium for green power). Participation rates in the programs with the 10 highest figures ranged from 3.0 to 5.8%.

Other entities throughout the world are also focusing more formally on the engagement of consumers and measurement of their participation in renewable energy programs. The National Institute for Research Advancement (NIRA) in Japan devoted an entire issue of its journal, *NIRA Policy Research*, to renewable energy development and citizen participation in the context of wind power.

Some of the most interesting and promising research concerning consumer attitudes and participation pertains to the new-home construction industry. Under the auspices of its Zero Energy Home (ZEH) program, DOE is promoting the construction of new single-family homes that combine available renewable energy technologies and energy-efficiency construction techniques to consume zero net energy from the utility grid. Both qualitative and quantitative research is currently being conducted at NREL on the consumer response to ZEH-like homes in San Diego, California. Although proprietary studies of homebuyers are relatively widespread in the building industry, this is some of the earliest work on customer response to ZEHs, particularly in a geographic locale in which volatile energy costs have been a political and economic concern.

The research involves direct interviews with purchasers of high-end homes (in excess of \$500,000) in a large development in which more than 250 highly energy-efficient homes with GPV may eventually be sold. Almost all the homes will have solar water heating as well. Initial interviews with the home buyers have targeted the following issues: reasons for purchase, preferences for standard or optional systems, views on energy efficiency and solar energy features, satisfaction with utility bills, aesthetics, construction and technological problem areas, impacts on energy-related behavior, homeowner knowledge, and overall homebuyer satisfaction.

Early findings were described in a paper presented at the 2002 Summer Study conference of the American Council for an Energy Efficient Economy. Among homeowners with GPV, the availability of solar features was frequently cited as one of the top three reasons for purchase. Some of these buyers said they were looking for a hedge against rising energy

costs and some were motivated by their own environmental values. Although indicators of satisfaction can be misleading in the early stages of technology deployment, all the individuals interviewed indicated they were satisfied and that they would make the purchase again. Their feedback, although expected to be representative of highly positive and proactive early adopters, provided key insights into the motivation of consumers to become involved with renewable energy technology and the marketability of ZEHs. The realization that so many of these homes would actually be sold is further evidence of the existence of consumer demand for renewable energy and renewable energy systems.

Both GPV and solar water heating are included in the Million Solar Roofs initiative directed by DOE's Federal Energy Management Program. The objective of the Million Solar Roofs program is to encourage and support the installation of 1 million solar energy systems on residential, commercial, and public sector buildings by 2010. If successful, the program will have had a major impact on public perception and renewable energy markets in the United States.

Solar water heating is a more mature technology than GPV, having been available in several forms for more than 20 years. Studies on the market response to solar water heating have been conducted by several entities. The actual size of the installation base is difficult to reliably estimate, but anecdotal estimates for the United States range from 1 to 2 million systems. In the international arena, the International Energy Agency conducted a 5-year program begun in 1998 to enhance the market for solar domestic hot water systems in Canada, Denmark, Germany, The Netherlands, and Sweden. With the ultimate objective of 10,000 installations, this program is similar to the Million Solar Roofs initiative in the United States. Results of the project have not yet been made publicly available. Based on a March 2002 survey, the Australian Bureau of Statistics estimates that in Northern Territory, approximately 53% of households have solar hot water systems, as do approximately 16% of households in Western Australia (although this number is down from its peak of approximately 20% in 1999). Although these percentages seem robust, they represent only approximately 4% of all households in Australia.

Finally, the 1998 study by the Energy Center of Wisconsin concerning wind power program participation reports that 24% of 647 survey respondents who were asked to make an actual commitment to pay more to receive wind power agreed to do so

(as previously noted, in another group of 761 individuals in the study that were asked, hypothetically, about their willingness to pay for green pricing, 43% responded affirmatively). When presented with a menu of monthly payment amounts, 47% committed to pay up to \$2 more per month or \$24 per year for wind power (compared to 61% of the other group who said, hypothetically, that they would be willing to pay up to \$2 more per month). Thirty-five percent committed to pay up to \$4 more per month (compared to 58% of the other group), 23% committed to pay up to \$8 more per month (compared to 50% of the other group), and 8% committed to pay up to \$24 per month (compared to 31% of the other group). A third group of 237 individuals were asked to pay for a fixed-size block of wind power of their choice every month, and surprisingly, more than one-third (37%) agreed to do so. The minimum block size was 50 kWh per month at an estimated incremental cost of \$2, and the maximum block size was 600 kWh per month at an estimated incremental cost of \$24. Other block sizes in multiples of 100 kWh were available at \$4 per 100 kWh per month.

As expected, the percentages of respondents in the Wisconsin study who actually agreed to sign up were lower than the corresponding percentages of respondents who were asked about their hypothetical willingness to pay. This is the only publicly available research that involves a direct comparison of actual consumer behavior regarding renewables to presumed, hypothetical behavior, and it is the only publicly available work that attempts to measure the difference. Nonetheless, it seems clear that consumers throughout the world are utilizing or purchasing renewable energy systems for personal use and that the trend toward greater acceptance and deployment is a positive, increasing one.

9. PUBLIC RESPONSE TO RENEWABLES IN THE TRANSPORTATION ENERGY SECTOR

The foregoing discussion focuses on electricity and power generation, particularly as they relate to the residential and commercial energy sectors. However, attitudes toward renewable resources carry over to other energy sectors as well. With respect to transportation energy, a decade-long emphasis on the development and proliferation of alternative fuels and

alternative fuel vehicles has produced consumer responses not unlike those exhibited toward electricity and power generation. Again, only limited information is available, and it must be noted that not all alternative fuels are derived from renewable resources.

Although the rate of growth since 2000 has slowed, the total number of alternative fuel vehicles in use has increased steadily in the past decade in response to direct support, mandates, and research and development activities of government agencies throughout the world. Whereas more alternative fuel vehicles have certainly been placed in service during this time, record keeping has also improved, so the estimated annual rates of increase, particularly in the early 1990s, may not be completely accurate. The EIA estimates that in 2002 there were approximately 500,000 alternative fuel vehicles in use in the United States, of which approximately 60% were privately owned and operated. Approximately 82,000 of these were estimated to be ethanol-fueled vehicles (E85 and E95 combined; E85 is an 85/15% mixture of ethanol and gasoline). In comparison, the EIA reported only approximately 45,000 alternative fuel vehicles in 1993, of which only approximately 450 were ethanol fueled.

Of all the fuels classified by the DOE as alternative fuels, only ethanol can be specifically classified as a renewable transportation fuel. Ethanol is derived mostly from corn. Methanol might also be regarded as having been derived from a renewable resource, depending on its composition. The number of methanol-fueled vehicles in use in the United States has declined rapidly since 1997 to fewer than 6000 in 2002, whereas the number of ethanol-fueled vehicles has increased. In fact, no new methanol-fueled vehicles are being manufactured. Among all types of alternative fuel vehicles, the number of ethanol-fueled vehicles is growing at one of the fastest rates, particularly in the Midwest, where corn is plentiful (Fig. 6).

Not all alternative fuel vehicles operate exclusively on alternative fuel (alternative fuel vehicles that operate exclusively on their target fuel are referred to as dedicated vehicles; those that do not are said to be nondedicated). Ethanol vehicles, for example, can typically operate on either ethanol or gasoline (these are often referred to as flexible fuel vehicles). Information about the percentage of time that such vehicles actually operate on ethanol, especially privately operated vehicles, is sketchy.

Despite the growth in alternative fuel vehicles, the total number is still only a fraction of the entire in-use vehicle population. The EIA estimates that in the

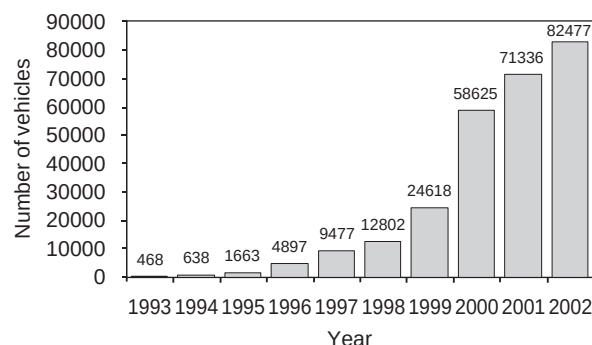


FIGURE 6 Annual numbers of ethanol-fueled vehicles (E85 and E95) combined. Totals for 2001 and 2002 are estimated. Data from the U.S. Energy Information Administration (2002b).

United States, alternative fuel vehicles represent only approximately 0.2% of total registered on-road vehicles. Furthermore, the amount of alternative fuel consumed annually in the transportation sector is less than 3% of total fuel consumption, and the percentage has declined since 1999. The EIA estimated U.S. alternative fuel consumption in 2001 to be 4.4 billion gasoline-equivalent gallons, of which ethanol accounted for 1.12 billion gasoline-equivalent gallons. A large percentage of ethanol is used as an additive in oxygenated gasoline formulations; thus, by default, many Americans—indeed many consumers throughout the world—are ethanol users.

A number of studies have been conducted regarding market penetration of alternative fuel vehicles. For example, in 1996 the EIA reported on the markets for alternative fuel vehicles in general, whereas an earlier study conducted at the University of California at Davis focused specifically on alcohol-fueled vehicles.

Three NREL reports from the mid-1990s provided synopses about the preferences and satisfaction of drivers of alternative fuel vehicles. The first one described the results of a survey of U.S. federal fleet drivers in which respondents were asked to compare the performance of alcohol-fueled vehicles to that of other types. Approximately 90% of respondents assigned overall ratings of “very good” or “excellent” to alcohol-fueled vehicles, and similarly high percentages of drivers indicated they were leaning toward being “satisfied” or “very satisfied” with their vehicles. Approximately 87% of drivers of ethanol-fueled vehicles and approximately 67% of drivers of methanol-fueled vehicles indicated they would recommend their vehicles to someone else. A second report described the results from a corresponding survey of federal fleet managers. A high percentage of

these individuals (approximately 70%) indicated that drivers of the alcohol vehicles in their fleets were neutral toward them (a somewhat less enthusiastic response than from the drivers), but more than 80% of the fleet managers indicated they received approximately the same number of complaints from drivers of alcohol-fueled vehicles as they did from drivers of gasoline-fueled vehicles. The final report described the results of a survey of drivers of state and city government fleet vehicles. More than 80% of the drivers of alcohol-fueled vehicles assigned them overall ratings of excellent or very good, and almost all drivers said they were about the same as gasoline-fueled vehicles. Approximately 70% of the drivers of alcohol-fueled vehicles said they were very satisfied with their vehicles, and approximately the same percentage said they would recommend their vehicles to other drivers.

A number of other surveys of fleet managers have been conducted, both by commercial entities and by professional research organizations. However, many such studies have statistical and operational shortcomings, and they are not specifically concerned with alcohol-based (renewable) fuels.

Detailed information regarding the preferences and perceptions of owners/drivers of privately owned alcohol-fueled vehicles is not available, although anecdotal evidence suggests similarly high levels of satisfaction. After all, individuals are unlikely to purchase these kinds of vehicles for private or business use if they do not already have a positive attitude toward them. Likewise, no recent data concerning willingness to pay for alcohol-fueled vehicles are available.

Like the consumer information concerning heat and power generated from renewables, this collective body of evidence suggests the public response to automobiles capable of operating on renewable fuels is positive. More expansive growth seems to hinge primarily on vehicle availability in the marketplace, accessibility to adequate fuel (adequate refueling infrastructure, ease of refueling, etc.), and both vehicle and fuel price.

Fuel cell technology promises to revolutionize automotive power systems and substantially reduce the environmental impact of transportation emissions. Fuel cells also show promise as power plants for other appliances and devices such as computers. Simply stated, fuel cells are devices that produce electrical energy from a fuel using an electrochemical reaction, or the process of oxidation, rather than combustion. The function of a fuel cell is similar to that of a battery. Although various fuels can be used, hydrogen

is the primary choice because the only by-products of the reaction are water, heat, and electricity. In a direct-hydrogen fuel cell vehicle, the controlled reaction between hydrogen and oxygen is believed to be more efficient than the process of internal combustion, in which the fuel (e.g., gasoline or diesel) is actually burned. Other fuels besides hydrogen can be used—particularly gasoline, diesel, ethanol, and methanol—but they all have to first be converted to hydrogen onboard in order for the fuel cell process to work. Safety concerns surrounding the transportation of hydrogen onboard private and commercial vehicles is the primary reason for considering other fuels, and there is considerable interest in renewable fuels such as ethanol and methanol because of their environmentally friendly characteristics. With support from various government agencies and other funding organizations, most of the automotive industry is engaged in fuel cell technology research. Honda and Toyota are the industry leaders; both companies have manufactured a limited number of fuel cell vehicles for sale (Toyota already has a sizable market, particularly in Japan, for its hybrid electric vehicles). Fuel cell systems for heavy vehicles and buses are also being researched by various organizations, such as Georgetown University and NREL. Although an important objective of many fuel cell vehicle research programs, there has been little investigation into public acceptance or consumer demand, and only limited information is available in the public arena. Consequently, only time will tell whether fuel cell technology can live up to its billing and supplant other power systems that are already deeply entrenched in the transportation sector.

10. FACTORS THAT MOTIVATE PUBLIC RESPONSE TO RENEWABLES

The question as to what motivates consumers to become personally involved with renewable energy is both interesting and multifaceted. The driving forces are myriad, but some of the more obvious ones are individualism and personal security, economics, environmental concern, the pursuit of a technocentric lifestyle, and even altruism. Although a direct linkage between such influences and actual purchase behavior is difficult to establish, survey data support the notion that individuals who are more favorable toward renewable energy also tend to be more sensitive to, or disposed toward, one or more of these factors.

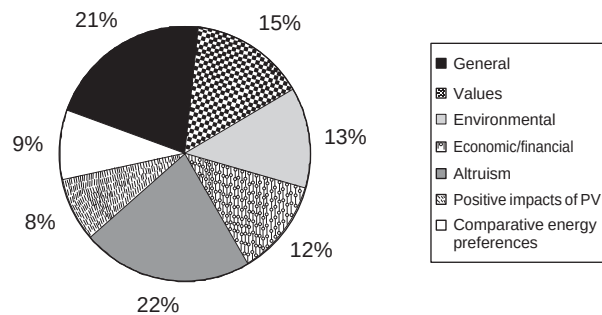


FIGURE 7 Percentages of references by participants in focused interviews to factors that motivate GPV adoption; $n = 3465$ total references by 120 participants in a purposive sample of residential energy customers in Colorado. Source: Farhar and Burhmann (1998).

A fairly extensive analysis of consumer motivation to adopt renewable energy technology is reported in the qualitative work by Farhar and Burhmann leading up to the 2000 NREL study of public response to GPV in Colorado. The percentages of all references in focused interviews to various motivating factors are summarized in Fig. 7.

The 2000 NREL study subsequently reports the percentages of respondents who indicated various perceived benefits of GPV to be important. These percentages are summarized in Table III. Taken together, the results of these two studies suggest that consumers' motivations to adopt renewable energy technology (e.g., GPV) are quite varied but perhaps more generally focused on economics and the environment.

The link between energy preferences and environmental concerns has been known for some time, but with increased emphasis on global warming and the greenhouse effect, the linkage is becoming more pronounced. National surveys conducted during the past 10 years indicate that fairly strong majorities of Americans believe there is a connection between energy usage and risk to the environment, and in the 2000 NREL study Colorado homeowners are shown to associate energy usage with environmental threat.

Although economic issues are always on the minds of consumers, they have perhaps been brought to the forefront by the recent energy shortages and price spikes in California. A 2001 brief prepared by Farhar on behalf of the Roper Center describes the results of several national polls conducted prior to the September 11, 2001, terrorist attack that indicate increasing concern among Americans about energy costs and the impact that worldwide shortages might have on them.

TABLE III

Percentages of Survey Respondents Who Indicated Several Perceived Benefits of Grid-Tied Photovoltaics (GPV) to Be Important

Perceived benefit of a GPV system	% of respondents indicating importance ^a
Result in long-term energy cost savings	86
Help reduce air pollution and acid rain in our area	84
Conserve natural resources	85
Benefit future generations	82
Reduce electricity bills right away	82
Help protect the environment	82
Increase resale value of home	79
Have electricity during a power outage	76
Help reduce global warming	75
Pay for itself over system lifetime	72
Sell excess electricity back to utility company	71
Increase self-sufficiency	70
Increase diversity of energy sources	70
Help create jobs for Colorado	68
Good for Colorado's economy	68
Encourages others to replace gas and coal use	62
Feels good to do this	51
Increase awareness of household energy use	46
An opportunity to make a difference in community	44
Help create and expand PV market	38
A new technology to enjoy	33
Profitable for utility companies	20
First on block to have a PV system	10

Source: Farhar and Coburn (2000).

^aOn a scale of 1 to 10, with 10 being "very important," the percentage of all respondents who chose 7–10.

Research conducted at the University of Michigan prior to the catastrophic events of 2001 suggests, however, that economics may not be the sole motivating factor. In 1998, Clark et al. noted that the motivation for individuals to participate in initiatives such as green power programs is more complex, encompassing both intrinsic and extrinsic factors. Recent research concerning the motivations for corporations and other nonresidential customers to become involved in green power complements the University of Michigan findings. A 2001 report

prepared by Lawrence Berkeley National Laboratory (LBNL) in conjunction with Ed Holt & Associates indicates that altruism is the most significant factor, particularly among smaller companies and firms. Although conventional wisdom might suggest that economic gain is the principal motivating factor, this work implies otherwise.

Taken together, this body of evidence suggests that economics, environmental concerns, and national security and self-sufficiency will be the forces driving consumer demand for renewable energy and renewable energy systems in the foreseeable future. This conclusion is not surprising, particularly to those who have broad knowledge of the issues facing the energy establishment. What is surprising is the rapidity and certainty with which the interaction of these forces is changing the worldwide political, industrial, and economic landscape. The wildcard in the proliferation of renewables may be the speed with which science, engineering, business, and industry can bring products to the marketplace to capture the imagination of a public eager to try new inventions that have the potential to reduce personal uncertainty.

11. MODELING PUBLIC RESPONSE TO RENEWABLES

Surprisingly, little has been done in the way of formally modeling or predicting the public response to renewable energy and renewable energy systems. The 1998 report by the Energy Center of Wisconsin pertaining to wind power program participation purports to develop predictive models, but no such work beyond descriptive percentage distributions is actually presented. The 2000 NREL report presents the results of an extensive multivariate analysis of the survey data aimed at characterizing the Colorado market for GPV, which in context can be used to predict consumer participation. Beyond these two endeavors, little has been done to specifically predict public participation or response. On the other hand, there have been many efforts to develop economic models and market penetration models pertaining to renewable energy and renewable transportation fuels, although some of this knowledge is becoming dated. A collaborative partnership among NREL, LBNL, and Ed Holt & Associates resulted in forecasts of growth in green power markets in the United States.

The problem is that modeling and predicting public response to any technological intervention is

difficult. Societal, cultural, economic, and other forces all interact to affect the responses of ordinary citizens. In addition, the identification and selection of appropriate measurements, and the collection of data in sufficient amounts and of sufficient quality, are substantial impediments to research. Finally, as previously suggested, behavior and attitudes are not necessarily directly related, and without a specific definition of what actually constitutes the public response, the ideas of modeling and prediction are necessarily reduced to purely theoretical exercises or to the reporting of mostly empirical, possibly indirect, observations.

12. CONCLUSION

Public attitudes toward renewable energy and renewable energy systems have shifted during the past 10–20 years, and they continue to evolve. The change is observable in ongoing opinion and market research conducted by government agencies, business entities, and respected research organizations throughout the world. Several factors are responsible. Most significant, perhaps, is a growing recognition that worldwide demand for energy—specifically, electricity—is surging and that traditional energy resources are limited. Recent economic and political instability, both globally and domestically, have created a milieu of uncertainty that accentuates public concern about the ability of government and industry to supply the energy necessary to meet societal needs. In addition, there is increasing evidence of global climate change, and consumers are becoming more convinced of its link to the consumption of traditional fossil fuels and more apprehensive about its impact on the environment and public health.

Nearly all the evidence of an attitude shift is empirical, and more needs to be done from an inferential and predictive standpoint. In addition, there are acknowledged issues of data quality and comparability. However, the amount of evidence has grown to the point that the positive trend toward renewables in public opinion and consumer attitudes cannot be denied. Operationalizing a favorable outlook into actual product purchases and program participation is a different matter, indeed; but here, too, there is substantial evidence that markets are developing and that consumers are increasingly acting on their convictions that renewable energy is a viable approach to sustainable economic and environmental security. As in the case of all technological ventures, price, availability of products,

accessibility to infrastructure, and reliability and performance of systems will be key to further consumer participation and expansion of markets.

SEE ALSO THE FOLLOWING ARTICLES

Public Reaction to Electricity Transmission Lines • Public Reaction to Energy, Overview • Public Reaction to Nuclear Power Siting and Disposal • Public Reaction to Offshore Oil • Renewable Energy and the City • Renewable Energy Policies and Barriers • Renewable Energy, Taxonomic Overview

Further Reading

- Adamson, K. A. (2003). An examination of consumer demand in the secondary niche market for fuel cell vehicles in Europe. *Int. J. Hydrogen Energy* 28(7), 771–780.
- Appelbaum, L., and Cuciti, P. (2003). “Electricity Generation, Renewable Energy and Environmental Considerations: Results of a Colorado Survey.” University of Colorado at Denver, the Wells Fargo Public Opinion Research Program, Denver, CO.
- Australian Bureau of Statistics (2002). “Environmental Issues: People’s Views and Practices,” ABS Catalogue No. 4602.0. Australian Bureau of Statistics, Canberra, Australia.
- Champ, P. A., and Bishop, R. C. (1998). “Wind-Power Program Participation: Developing Predictive Models,” Report 179-1. Energy Center of Wisconsin, Madison, WI.
- Farhar, B. C. (1993). “Trends in public perceptions and preferences on energy and environmental policy,” NREL/TP-461-4857. National Renewable Energy Laboratory, Golden, CO.
- Farhar, B. C. (1996). *Energy and the environment: The public view*, Issue Brief No. 3. University of Maryland, Renewable Energy Policy Project, College Park, MD. [Available: <http://www.repp.org>]
- Farhar, B. C., (1999). “Willingness to pay for renewable electricity: A review of utility market research,” TP-550-26148. National Renewable Energy Laboratory, Golden, CO.
- Farhar, B. C., and Burhmann, J. (1998). “Public response to residential grid-tied PV systems in Colorado: A qualitative market assessment,” TP-550-24004. National Renewable Energy Laboratory, Golden, CO.
- Farhar, B. C., and Coburn, T. C. (2000). “Colorado homeowner preferences on energy and environmental policy,” NREL/TP-550-25285. National Renewable Energy Laboratory, Golden, CO.
- Farhar, B. C., and Houston, A. H. (1996). “Willingness to pay for electricity from renewable energy,” NREL/TP-260-21216. National Renewable Energy Laboratory, Golden, CO.
- Farhar, B. C., *et al.* (2002). Market response to new zero energy homes in San Diego, California. In “Proceedings of the 2002 ACEEE Summer Study.” American Council for an Energy Efficient Economy, Asilomar, CA.
- Ferguson, E. G. (1999). “Renewable Resources and Conservation: what consumers want.” Bonneville Power Administration, Portland, OR.
- Gielecki, M., *et al.* (2001). Incentives, mandates, and government programs for promoting renewable energy. In “Renewable Energy 2000: Issues and Trends,” DOE/EIA-0628. U.S. Energy Information Administration, Office of Coal, Nuclear, Electric, and Alternative Fuels, Washington, DC. [Available: <http://www.eia.doe.gov/cneaf/solar.renewables.rea.issues/reaissues.sum.html>]
- Rauluk, V. (2001). “Development of markets for solar water heating systems,” Venture Catalyst Inc. final report 1998–2001 prepared for the U.S. Department of Energy, DOE/EE10357-F. U.S. Department of Energy, Washington, DC.
- U.S. Energy Information Administration (2002a). “Alternatives to Traditional Transportation Fuels 2000.” U.S. Energy Information Administration, Washington, DC. [Available: <http://www.eia.doe.gov/cnea/alternate/page/datatables/atf13.00.html>]
- U.S. Energy Information Administration (2002b). *Renewable energy annual 2001, with preliminary data for 2001*, DOE/EIA-0603. U.S. Energy Information Administration, Office of Coal, Nuclear, Electric, and Alternative Fuels, Washington, DC. [Available: http://www.eia.doe.gov/cneaf/solar.renewables/page/rea_data/rea.sum.html]
- Whalen, M., and Coburn, T. (1997). “Perspectives on AFVs: 1996 federal fleet driver survey,” NREL/BK-540-22721. National Renewable Energy Laboratory, Golden, CO.
- Wiser, R. H., *et al.* (2001a). “Public goods and private interests: Understanding non-residential demand for green power,” LBNL-47300. Lawrence Berkeley National Laboratory, Berkeley, CA.
- Wiser, R. H., *et al.* (2001b). “Forecasting the growth of green power markets in the United States,” NREL/TP-620-29408. National Renewable Energy Laboratory, Golden, CO.

Radiation, Risks and Health Impacts of

RICHARD WILSON

Harvard University
Cambridge, Massachusetts, United States



-
1. Introduction
 2. Effects at High Doses (> 3 Sv)
 3. Effects at Moderate to High Doses (0.3–3 Sv)
 4. The Effects of Dose Rate
 5. Effects at Low and Very Low Doses
 6. Weighting Factors (W_r), or Relative Biological Effectiveness
 7. Policy Implications
 8. Public Perception

Glossary

absolute risk model A model in which the increase in the rate of cancer, or other disease, is independent of the natural incidence at that age and place.

absorbed dose The amount of energy absorbed in a material. The Système International unit for absorbed dose is named after the medical physicist Professor L. H. Gray. The gray (Gy) represents an energy absorption of 1 joule per kilogram of absorbing material, i.e., $1 \text{ Gy} = 1 \text{ J/kg}$. However, an older unit, the rad, is still often used, particularly in the United States ($100 \text{ rad} = 1 \text{ Gy}$).

ALARA An acronym for “as low as reasonably achievable,” a criterion for judging radiation protection; alternatively, ALAP (as low as practicable) or ALARP (as low as reasonably practicable).

assigned share See **probability of causation**.

equivalent dose (radiation weighting factor) An expression introduced to take into account the dependence of harmful biological radiation effects on the type of radiation being absorbed. The same absorbed dose delivered by different types of radiation may result in different degrees of biological damage to body tissues. The total energy deposited is not the only factor that determines the extent of the damage. The equivalent dose is therefore a measure of the risk associated with an exposure to ionizing radiation. Risks due to exposures to different radiation types can be directly

compared in terms of equivalent dose. The unit of equivalent dose is the sievert (Sv), defined for a given type of radiation by the relationship $H = D W_r$, where H is the equivalent dose in sieverts, D is the absorbed dose in grays, and W_r is a numerical weighting factor. In older centimeter-gram-second (CGS) units, the term “roentgen equivalent man” (rem) is still often used ($1 \text{ Sv} = 100 \text{ rem}$).

excess relative risk (ERR) The excess (additional) risk caused by radiation, divided by natural risk (incidence).

ionizing radiation The radiation with individual quanta of high enough energy to ionize an atom.

LD₅₀ The “least dose” at which 50% or more of people are affected (usually producing, but not always, death).

probability of causation The estimated probability that an event (such as cancer) was caused by a postulated action (usually radiation exposure).

radiogenic cancers Occurrence of cancers at a body site where it is known that radiation can cause an increase; most leukemias are radiogenic cancers, whereas chronic lymphocytic leukemia is not.

relative biological effectiveness (RBE) An alternative description for the weighting factor W_r .

relative risk model Extrapolation of the ratio of a risk to the risk from natural causes at a certain age and place; this model is in contrast to the absolute risk model.

Each and every one of us is exposed to natural ionizing radiation, all the time, from the sky, from the air, from the ground, and even from our own bodies. Mankind can also produce ionizing radiation. This was realized when X rays were discovered by Professor Roentgen in 1895; radioactivity was discovered by Becquerel soon thereafter, and scientists began to notice that exposure to radiation had harmful effects—radiation was noted to burn, damage, and even destroy tissue. These effects were used very soon (about 1905) for beneficial purposes, to destroy an unwanted skin tumor. Almost from the

beginning arose the dilemma: excessive radiation can be bad for you, but it can also be good for you. The focus here is on assessment of the risks and the benefits of radiation.

1. INTRODUCTION

The most important medical application of radiation is the use of X rays for diagnostic purposes. There is an enormous advantage to a physician, and hence to the patient, of knowing exactly where a bone or blood vessel lies in the body. Early on, this realization seemed to outweigh any small risk that continued use of X rays might entail. This crude risk–benefit comparison probably contributed to the fact that it was many years before people acted to reduce radiation risk. The benefit seemed so great that it was forgotten that risk could be reduced without changing the benefit. In the 1920s and 1930s, a number of pioneer radiologists suffered severe injuries, and some even died, as a result of prolonged exposure to dangerously high intensities of X rays. These early workers in the field had no means of accurately measuring the harm caused by radiation and they often used inferior measurement equipment even though better equipment was available; to establish radiation exposure, they depended on unreliable effects, such as the degree of skin reddening caused by the exposure, or on timing the exposure from a certain type of X ray machine.

Other adverse effects of radiation were noted in workers involved in several industrial or clinical uses of radiation in the first half of the 20th century. In the 1920s, watch dials were painted with a luminous paint containing an alpha particle emitter, usually from radium. Most watch factory painters sucked the paintbrush to obtain a fine point, thereby absorbing radium in their mouth, ultimately in their bones. In clinics and hospitals, physicians suffered from excess radiation when they administered large doses of radiation to cure painful diseases such as ankylosing spondylitis. X-Ray operators failed to shield themselves and the parts of the patient's body that were not under study. Radiologists would sometimes look straight at a fluorescent screen and therefore straight into the X-ray tube.

The radiation dose for a typical chest X ray was about 1 rem until about 1960; with good shielding and use of image intensifiers it can now be (and could have been 50 years before) 7 mrem. The big impetus toward understanding the health effects of radiation

was the massive expansion of nuclear technology following the development of nuclear reactors, and the explosion of the atomic bombs in 1945. Since then, radiation has become inextricably linked with the atomic bomb and its adverse effects, much to the detriment of benign uses. The fact that radiation is associated with bombs leads many people to believe that radiation is uniquely dangerous. That is far from the case. For example, the largest and most important study of radiation-induced cancer comes from the study of Hiroshima and Nagasaki victims, among whom 500 or so excess cancers (more than expected from the nonexposed population) were observed. Yet that many excess lung cancers were found in a separate study unrelated to exposure from the bombs. One crucial feature of such data is that radiation does not seem to cause any medical or biological effect that cannot also occur naturally. It merely increases the probability of the effect. This fact is very important, both for understanding possible dose–response curves and for deciding what, if anything, to do about radiation at low doses. It also leads us to ask a general question: Does radiation add an additional probability of developing cancer (an absolute risk model) or does it multiply the probability that is there from natural or other causes (a relative risk model)? Although both models are discussed in most summary reports, it is noteworthy that the emphasis has changed from an absolute risk model (in 1970) to a relative risk model with an increase in calculated risk—particularly in aging population groups.

2. EFFECTS AT HIGH DOSES ($> 3 Sv$)

When radiation is given over a short period (less than a day) at doses greater than 200 rem, the primary effect is serious destruction of blood and tissue. Two situations can be distinguished: radiation exposure of a small area of the body and exposure of a large percentage of the body surface. A recent example of the first situation arose in the republic of Georgia, of the former Soviet Union, when a soldier picked up a large strontium 90 source that had been used for a thermoelectric generator and left behind by the Soviet army. The soldier's hand was treated for a year in the hospital of the Curie Institute in Paris before the soldier fully recovered.

In the past, misuse of X rays led to many situations in which a hand or other extremity was

heavily exposed (up to 200 Sv = 20,000 rem). At this level, tissue is destroyed but can be surgically replaced. High exposures of limited quantities of tissue occur in radiotherapy. The aim in radiotherapy is to provide a large enough dose to destroy a tumor, without destroying neighboring tissue. Typical doses to the tumor are 60 Sv (= 6000 rem). If the dose is given in several fractions, the neighboring healthy tissue can recover. In contrast, a much smaller dose given to the whole body can destroy the whole organism. Ordinarily, the dose from which half of the recipients will die, the LD₅₀, is about 2.5 Sv (= 250 rem), but the LD₅₀ can be extended up to about 5 Sv in aggressive treatment. Following high-dose radiation, blood transfusions to bring in healthy blood have been successful, and bone marrow transplants to replace destroyed bone marrow were tried in a few cases of heavily exposed workers after the 1986 Chernobyl nuclear accident. A couple of the Chernobyl patients improved, but an equal number were probably killed by the treatment. At 5 Sv (= 500 rem), the gut and many other body organs begin to fail. Death is not immediate but takes up to a week. One famous victim, the Canadian Dr. Louis Slotin, a scientist at the Omega Site Laboratory in Los Alamos, New Mexico, lived for a month, and actually wrote a description of the incident while he was dying, after an accident with supercritical plutonium in 1946 exposed him to ~10 Gy of gamma and neutron radiation.

3. EFFECTS AT MODERATE TO HIGH DOSES (0.3–3 Sv)

3.1 Chronic vs. Acute Effects

Daily doses lower than those required to destroy tissue, can, if repeated over a long period, cause health effects. In contrast to the acute effects from exposure over a short exposure, these chronic effects often involve whole body doses accumulated over years, of 10 Sv and more. From general biological principles, the following effects of continued exposure to even moderate doses of radiation might be expected: induction or aggravation of cancer, induction or aggravation of heart disease, induction or aggravation of other diseases, production of genetic anomalies that can be passed on to another generation, and induction of birth defects by genetic mutation. Most of the studies address only cancer induction, and the data do indeed suggest that cancer induction is the dominant effect. The data are better

for cancer than for the other outcomes largely because the observed effects are greater. Several groups of radiation-induced cancer (e.g., leukemia, solid cancers, lung cancer) can be distinguished based on their different characteristics.

3.2 Chronic Radiation Sickness

Normally, any worker who receives a high radiation dose on the job is prohibited from participating further in any work that involves radiation exposure. A case history in which this practice was ignored occurred in the late 1950s, in the former Soviet Union, where there was a great push to build an atomic bomb to “balance power” with the United States. For many years, it had been rumored that workers at the Mayak nuclear reprocessing plant in the southern Urals received large radiation doses. Attempts from the western countries to discover the facts about what happened at Mayak were fruitless until 1991, when a description was published in Russian in the journal *Priroda*. The workers at Mayak had continued working even after being highly exposed, and over a 2-year period received doses up to 10 Sv (= 1000 rem). The first diagnosed effect was a drop in blood cell count and general sickness. This was called chronic radiation sickness, in contrast to acute radiation sickness, which is caused by large singular doses. Radionuclides from the Mayak chemical processing facility were dumped into neighboring Lake Karachay. In 1955 and 1956, water overtopped the reservoir and spilled into the Techa River, which drains into the Arctic Ocean. Villagers drank the water and ingested many radionuclides. For 40 years, the health of 30,000 villagers around the Techa River has been studied and chronic radiation sickness has been diagnosed. Around the Techa River, the exposures were mostly internal; the internal organ dose is used in describing leukemia incidence. In contrast, “solid” cancer diagnoses depend on external doses, which are far less easily determined. The internal Techa River exposure determination was based on detection of the “bone seeker” strontium-90; doses were fairly accurately determined by examination of radioactivity of teeth and other bones.

3.3 Leukemia

Although there have been studies of the effects of radiation on people for 100 years, the most important studies are the studies of the consequences of the Hiroshima and Nagasaki atomic bombs. The

research on the survivors has involved many good scientists, considerable effort, and considerable expense, mostly at the Radiation Effects Research Foundation in Hiroshima. The exposures occurred many decades ago, but an increase in cancers over that expected in the general population is still occurring. Therefore, the most recent research is of greatest importance. In addition, the radiation dose to which the population was exposed is uncertain. It was derived from measurements at other (similar) nuclear explosions, and based on the neutron dose by measuring long-lived radioactivity in the region.

Leukemia was the first cancer to be noticed in the atomic bomb survivors. Leukemias began to appear 5 years after the exposure and continued to appear for another 20 years, after which the incidence of leukemias (almost) ceased. Radiation induces leukemias more readily than other cancers. Therefore, leukemias are often considered to be a “marker” for radiation effects. But the variation with age is clearly in great contrast to that of the “solid” cancers, and in the elderly even an “absolute risk” model would overpredict the number.

3.4 Solid Cancers (Excluding Lung Cancer)

Solid cancers occur in tissues of the body and are not connected with blood components. They do not appear until 20 years after radiation exposure and have been noted to continue to appear for 50 years. The increase of cancer with radiation dose seems to follow a “relative risk” pattern whereby the risk, after the latent period of 20 years, is increased by a constant fraction relative to the background at that age.

3.5 Lung Cancers

Lung cancers are of particular interest because the lung is the point of entry for a major radiation exposure. In order to understand the effect on humans of alpha particle irradiation, there have been several animal studies that show an increased incidence of cancer related to radiation dose. The animal studies showed a quadratic relationship. Therefore researchers initially looked for a quadratic relationship, and not a linear relationship, in human lung cancers. Several cohorts have provided useful information.

One of the surprises of the atomic age (i.e., since 1945) was that an excess of lung cancers appeared among uranium miners. In some of the early studies,

lung cancers seemed to appear only among Caucasian miners and not among Native American Indians. There was some speculation of a racial difference in sensitivity. This speculation was proved incorrect both by follow up and by observations that there were fewer tobacco smokers among the Indian miners, compared to the Caucasian miners. Most analysts have assumed that lung cancers are due to exposure to radon gas in the mines rather than to any other pollution in the mines (dust, etc.), although there is still room to question this. The main concern is that inhaled particles or radionuclides might cause lung cancer. The generally accepted effect of radon gas is that radon alpha particles and “daughter” products irradiate the lung and induce cancer. This suggests that exposures to other alpha particle emitters also should produce lung cancers.

There have been several studies of the health of persons who worked with, and possibly inhaled, plutonium. Lung cancer is the principal effect that may be caused by plutonium inhalation. Although there have been many persons working with plutonium, in Los Alamos (New Mexico), Rocky Flats (Colorado), and Hanford (Washington), safety precautions have kept exposures small except in one cohort, the workers at Mayak (the former Soviet Union) in the 1950s. Because plutonium deposits in the bone, it was possible in most cases at Mayak to make accurate estimates of exposure by looking at the deposited radioactivity in the bone, either *in vivo*, or after autopsy. The highly exposed group of Mayak workers developed lung cancer at twice the rate of the less exposed workers, whose lung cancers are assumed to be caused by cigarette smoking. The first studies found agreement with the expected quadratic variation of incidence with dose, but later studies showed that a linear relationship is more probable.

There have also been “ecological” studies of the relationship between the average lung cancer incidence in a population and the average concentration of radon gas. In the most complete of these studies, radon gas was measured in half a million U.S. households and data were compared to the average lung cancer incidence in the region. To the surprise of the author of the studies, the lung cancer rate was found to be lower in high-radon areas compared to low-radon areas. The effect is highly significant. There is argument among experts as to whether these data indeed show that inhaled radon gas is beneficial. On one side, modern epidemiologists argue that such “ecological” studies show little by themselves. All they show is that the average lung cancer incidence rate $\langle i \rangle = -k * \langle c \rangle \times \langle p \rangle$, where $\langle c \rangle$ and $\langle p \rangle$

are the average concentration and the average population density. They do not show that $c = k * a * b$. Moreover, there is, in their view, too much room for bias. In particular, the study includes both smokers and nonsmokers intermingled, and the largest cause of lung cancer is known to be cigarette smoking. On the other side, scientists note that data are data; the data do not fit the regulatory model proposed by the U.S. Environmental Agency (EPA), and that model must be incorrect or at least inadequate. The general consensus is that, at the concentration levels of concern, radon gas cannot be the major contributor to lung cancer, either for smokers or nonsmokers.

3.6 Childhood Cancers

In the period from 1940 to 1970, it was common to give an X ray to pregnant women to discover any problems with the infant fetus. This dosed the fetus with radiation approaching 0.01 Sv (= 1 rem). In a classic study (the Oxford study), it was found that the probability of developing childhood leukemia in the age group 0–8 years increased with the number of X rays. This effect was not observed in the Hiroshima–Nagasaki studies. It is usually accepted that these leukemias are caused by something that happened *in utero*. But there is a possibility that causality proceeds in the opposite direction. A physician noticing an abnormality leading to cancer might call for an X ray. The correlation might then merely be between the abnormality and a subsequent cancer. Childhood cancers occur up to 7 years from birth, so cancers from the Chernobyl accident should have been seen. A small effect is possible; it would not be larger than, and could be smaller than, the effect suggested by the Oxford study. It would be especially interesting to examine this more carefully by looking at all children of the women who were pregnant and who were exposed at Chernobyl or any other radiation accident.

3.7 Thyroid Cancers

In any nuclear reactor accident, one of the first determinations must be whether any radioactive iodine was released. The short-lived iodine isotope can deposit in pastures; if cows ingest contaminated pasture grass, the iodine will concentrate in their milk. If children drink the milk, the iodine will concentrate in the thyroid, resulting in a large radiation dose, up to 100,000 times normal. This happened at an accident in Windscale, U.K. in 1956, and at Chernobyl

in 1986. The first preventive measure must be to break the route of contamination by stopping all consumption of milk from the affected area. This was done at Windscale, but not at Chernobyl because of a desire to maintain secrecy. Many western scientists brought this to the attention of the scientists and politicians in the former Soviet Union, but when thyroid cancers were diagnosed among children starting in 1979, it was a great surprise.

Thyroid cancers, particularly among children, are usually not fatal. They are closely linked with benign thyroid nodules, so it is hard to get a diagnosis that is consistent across many groups. Initially (in 1989 and 1991), therefore, an increase in children's thyroid cancers in the former Soviet Union was attributed to a sudden increase in attention paid to the problem by physicians. However, this was soon shown to be incorrect. By 2002, over 1500 childhood thyroid cancers were seen in Belarus, Ukraine, and Russia, with excess relative risk of over 50. Fortunately, there has been considerable attention paid to these by western physicians, and only 11 fatalities have occurred so far. Increases in childhood thyroid cancers have not been seen elsewhere. Now that many years have passed since the Chernobyl accident, an increase in adult childhood cancers might be expected. Indeed, an increase has been seen, but such increases are hard to interpret, because of the twofold effect of increased attention to the problem by physicians and the fact that it is hard to develop consistent diagnostic criteria. One Russian study claims to account for the difficulties by searching for a dose-related effect, but none was found. Nonetheless, whether adult thyroid cancers have truly increased since 1979 (when an apparent increase was observed) remains a matter of controversy.

3.8 Heart Disease and Other Chronic Outcomes

A careful examination of the Hiroshima–Nagasaki data shows an increase in several other medical outcomes. In particular, heart disease appears with a frequency of about one-third the frequency of cancer.

3.9 Genetic Effects

There is a general belief, based both on general biological principles and on animal experiments, that radiation can cause genetic defects, but no well-authenticated examples attributable to radiation have been observed in humans, even in the Japanese

atomic bomb survivors. Again, the numerical value of the dose response is controversial. However, for both cancer and genetic defects, the disagreement is over very small numbers, i.e., whether the effects around 1 mSv (= 100 mrem) are very small or very, very small—much smaller per capita compared to the cancer effects.

3.10 Growth and Development of Children

A statistically significant adverse effect of radiation has been found in the Radiation Effects Research Foundation (RERF) data on children whose parents were irradiated while they were *in utero*. Those whose parents were exposed within 1500 m of the hypocenter of the Hiroshima bomb were on average 2.25 cm shorter, 3 kg lighter, and 1.1 cm smaller in head circumference compared to those exposed further away. Although (just) statistically significant, this effect needs confirmation in other cohorts before it can be generally accepted. This has not been confirmed in any other cohort. However, although other exposed groups exist, such as those exposed near Chernobyl or at the Techa River, there has been no systematic study that either confirms or denies the relationship. Another effect that is perhaps associated with (or a consequence of) the reduction in head size was seen in the RERF data. There is a statistically significant reduction in Intelligence Quotient (IQ) among children exposed *in utero*. In other cases, there is severe mental retardation. Although these effects on children are generally believed to have a threshold, the data on reduction in IQ are also consistent with a linear relationship between reduction and dose.

4. THE EFFECTS OF DOSE RATE

The previous discussion of the cancers caused by radiation exposure implies that it is the accumulated dose that is important. This seems to be a good first approximation. But from general principles one might guess that a high radiation dose given at a low rate over a period of years might have a different (probably smaller) effect than the same dose given in a short time, although if the time period is long, on the order of a lifetime, the difference is likely to be small. Data from animal exposures to radiation show that there is a reduction in cancers (for the same total dose) at low dose rates. A dose rate reduction factor

(DRRF) is usually introduced to describe this. From animal data, a DRRF of between 2 and 10 is found and can be used for radiation protection purposes. Situations of accumulated radiation doses of 0.5 Sv (= 50 rem) exist, including the Mayak and Techa River cohorts noted earlier and the population exposed to Cs-137 subsequent to the Chrenobyl accident. From the Techa River and the Mayak data it seems that the DRRF for leukemia is about 3, but with a large error band (perhaps from 2 to 6). The DRRF for solid cancers is about 1, with a much larger error band ranging up to a factor of 10.

The importance of the data from these accident situations in which large cumulative doses, but not large acute doses, were received can not be overestimated, as they are useful for predicting the effect of nuclear accidents, as distinct from nuclear explosions. However, the data are not accurate enough to be used directly, and so at the present time, the world must depend primarily on the Hiroshima-Nagasaki data with an associated DRRF.

5. EFFECTS AT LOW AND VERY LOW DOSES

5.1 Early Belief in Beneficial Effects

Soon after natural radioactivity was discovered in the 1890s, there arose a general belief that exposure to moderate levels of radiation was therapeutically beneficial. This belief was not unique to radiation exposure, and many known poisons were given for therapeutic purposes. The small amounts of arsenic in Fowler's solution cured many disorders, and arsenic was also used in larger doses to cure syphilis before other, safer antibiotics were available. Communities with mineral hot springs added "radium" to their names to attract customers. The name Radium Hot Springs appeared in many places. As late as 1980, the radioactivity level of some European mineral waters was listed in advertising as a positive feature. However, the data supporting the suggested beneficial effect are scanty. Although there were several reports of cures, it is unclear whether these cures were due to radium, to other minerals in the water, or merely to the beneficial effects of a long soak in hot water.

Some old experiments on animals suggested that moderate radiation doses led to a reduction in disease and to a prolongation of life. But these reports have not been replicated with recent protocols, nor have there been human studies with modern

epidemiological methods. It is completely consistent that radiation would cause a beneficial effect for one medical outcome and an adverse effect for another one. It seems very plausible that radiation could reduce infectious disease by destroying the germs. But in the last 100 years, infectious diseases, with the exception of autoimmune disease syndrome (AIDS), have been controlled by other methods, so that even if it is true that radiation reduces the effect of infectious disease, it may now be irrelevant. It may be relevant for AIDS, but no study has been made. Alcohol in high doses is well known to cause cancer and has a narcotic effect at still higher doses. But at low doses—up to one drink a day—alcohol reduces the risk of stroke.

Since 1945, there has been a search for some chemical that can reduce the effects of radiation. A few substances can cause excretion of undesired radionuclides. It has been suggested that the cellular molecule MDC1 (mediator of DNA damage checkpoint protein 1) acts like an emergency service in cells to detect and repair damage to DNA caused by radiation. It is too early to say whether the effect is significant.

Although it is generally assumed that radiation effects at any site or organ will follow similar dose-response relationships, this is not necessarily true. The National Toxicology Program database lists a couple of dozen chemicals that decrease the likelihood of cancer in rats or mice when applied at one site, but increase it when applied at another.

5.2 The Definition of Low Dose

The concept of what constitutes a low dose has been modified considerably over the past 50 years. In the first half of the 20th century, physicians believed that 0.01 Sv (= 1 rem) was a low dose. In 1945, a typical chest X ray administered a dose of 0.01 Sv and at least one jurisdiction (United Kingdom) went as far as to propose mandating such a X ray every year to help fight tuberculosis. Most experts now believe that this would have caused an excess number of cancers. The proposed U.K. bill died in the House of Lords because of the objections of the physicist Lord Cherwell. A mandate for a yearly chest X ray would now be more reasonable. Chest X rays are now routinely performed with a dose 150 times smaller. Also, a skin test for tuberculosis antibodies now makes a chest X ray unnecessary in most cases.

The idea that 0.01 Sv per year is a low dose has changed. In 1987, the U.S. Nuclear Regulatory Commission proposed to label “Below Regulatory

Concern” a radiation dose exposure of no more than 1 mrem (0.00001 Sv) to any person. The proposal was withdrawn after some vocal public opposition. This dose is 1000 times smaller than what was medically accepted before 1945! Moreover, natural background exposures are a few hundred millirems, or 100 times this amount, and vary considerably from place to place and person to person. For scientific and political reasons, “low dose” now means doses as low as, and usually well below, background doses. But scientifically there is still no generally accepted evidence for an effect, beneficial or adverse, of a radiation dose this small for any medical outcome

5.3 Dose-Response Curves of Acute and Chronic Radiation Exposure

Acute effects seem to occur only with radiation doses above a definite threshold dose. They are sometimes called deterministic effects. When a population is exposed and the dose is below that threshold, no one is hurt; everyone is hurt when exposure is above that threshold. This threshold dose seems to be independent of gender, genetic predisposition, and of racial characteristics. Even skin effects, which might be expected to be affected by pigmentation, seem to be independent of race. There are small variations and a dose-response curve is usually drawn as in Fig. 1. The dose at which 50% of persons in a population are affected is called the LD₅₀. For whole body exposure, the LD₅₀ is about 2.5 Sv (= 250 rem) and persons exposed at half this dose survive. At a dose that is one-fourth of the LD₅₀, it is generally accepted that acute effects will not occur.

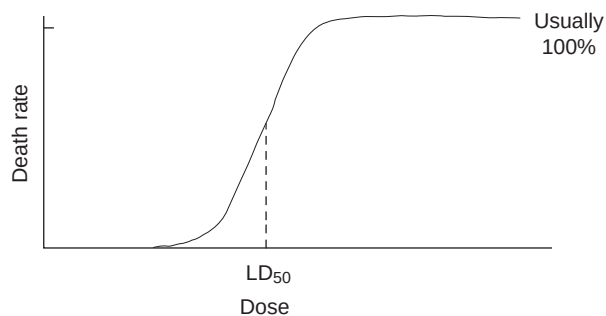


FIGURE 1 Dose-response curve, graphically representing the acute effect of a lethal dose of radiation. The LD₅₀ is the dose at which 50% of persons in a population are affected. A single acute dose or an accumulated acute level of radiation over a short time causes death, whereas one-tenth the lethal dose, repeated 10 times, does not cause death.

In contrast, the chronic radiation effect of cancer seems to follow a different dose-response curve. For those dosed at the LD_{50} , cancer seems unimportant. Even of those surviving after an exposure at the LD_{50} , only about 10–20% will develop cancer. However, in contrast to acute effects, at a quarter of this dose (0.5 Sv), many persons will develop cancer. Moreover, no one has been able to determine which persons in an exposed group will develop cancer and which will not. The dose-response relationship follows a flatter curve. Various possibilities to extrapolate high doses from the data are shown in Fig. 2. These include the linear; the superlinear, in which the effect at low doses is higher than the linear; the threshold, with zero below the threshold; and one with hormesis (from the Greek, *hormaein*, to excite), which is a beneficial effect below the threshold. That a linear curve is theoretically possible was noted in 1922 by the physicist Geoffrey Crowther, who produced the first theory of radiation-induced cancer. The idea is that when a cell is ionized by radiation it will initiate a cancer. The probability of ionizing a cell in a given time is clearly proportional to the radiation intensity and hence affords a theory that cancer induction is linear with dose even at low doses. But this theory in its simplest form cannot be true. Cosmic rays and background radiation ionize millions of cells every day and yet lifetime cancer incidence is only about 30% in the U.S. population. Other effects must modify this idea drastically. Cells can be repaired; they can be excreted without leading to a cancer, and so on. Whether cancer incidence is linear with dose depends, therefore, on whether these important mechanisms vary with dose. A discussion of these

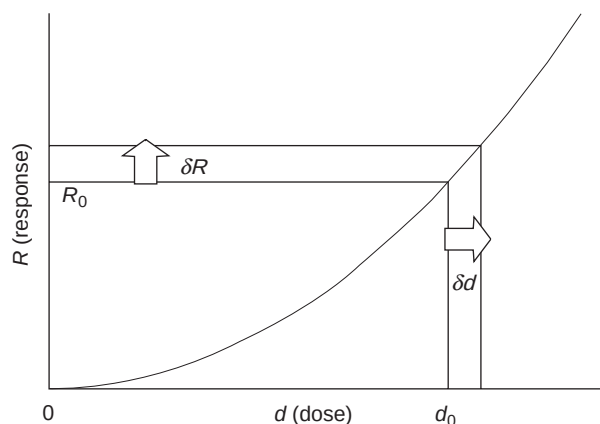


FIGURE 2 Threshold dose-response relationship ($R = Ad^3$), with axes shifted to R_0 and d_0 . Note that δR_0 is proportional to δd_0 .

becomes very complex. Nonetheless, it is therefore generally believed that the development of cancer, whether induced by radiation or other agents, natural or anthropogenic, is basically a stochastic process.

The discussion in the preceding paragraph does not, by itself, tell us the behavior of the dose-response curve at background levels. But it raises the possibility that it is a linear dose-response, down to zero additional dose. This is the assumption that is now usually made for all cancers.

5.4 Variation of Cancer Incidence with Background Exposure

Mankind is exposed to a variety of radiation from naturally occurring sources, often called background exposures. These background exposures have been extensively measured and vary widely over the globe. The average North American or Canadian receives about 0.4 mSv (= 40 mrem) per year from internal radioactivity and cosmic rays, 0.6 mSv (= 60 mrem) per year from medical and dental diagnostics, and about 0.02 mSv per year from other miscellaneous man-made sources. Among these sources, nuclear power stations contribute less than 0.001 mSv. Collectively, background radiation exposures total about 1 mSv (= 100 mrem) per year, although persons living at high altitude have a higher radiation exposure than do those at sea level. Many studies have therefore attempted to relate health indices to altitude to deduce the effect of radiation. In many studies, cancer mortality seems to be lower in the high-altitude areas. In particular, the radiation levels in the Rocky Mountain states are higher than they are in the Gulf Coast states, yet the cancer rate is lower. This effect may be seen throughout the United States and Canada, but confounding factors may exist. Many Mormons, who do not smoke or drink alcohol or coffee, live in Utah and mountain states and seem to have half the cancer rate of their non-Mormon neighbors. Many analysts conclude that the only fact of importance from studies of geographical variation is that radiation at levels of about 2–3 mSv per year, or 0.1–0.2 Sv (10–20 rem) per lifetime, is not an important factor in developing human cancers compared to other factors.

In addition to high altitudes, there are other areas of high radiation background in the world. High levels of ambient radiation are encountered in Brazil, China, and India, for example. A population of several hundred thousand people living in coastal

areas of Kerala, South India, is exposed to natural radiation levels up to 70 mSv (= 70 rem) per year, with half the population exposed to doses higher than 2.6 mSv (= 260 mrem) per year. The observed incidence of cancer is about three-quarters of that in North America. This cannot be discounted as a consequence of early deaths due to other causes, because the population's life expectancy is 74 years for women and 69 years for men—only a few years less than life expectancy in North America. There has been no observation of an elevated incidence of malformations in newborns in this population. Although most experts are unwilling to accept these data as definitive evidence for a beneficial effect at low doses of radiation, these studies do show unequivocally that ordinary people who do not worry about the geographical distribution of cancer rates, which is almost everyone, should not worry about radiation at these doses.

5.5 General Argument about Effects of Low Doses

When radiation doses of concern are considerably greater than background doses, it is possible to discuss logically the effects of radiation, independently of whatever causes the background. In discussing radiation doses at levels comparable to the background, or even 100 times smaller, it is necessary to consider “introduced” and background doses together. However, very few scientists and scientific studies do this logically. The most important fact in such a discussion is that introduced radiation does not seem to induce any cancers that are not seen in the absence of radiation and are therefore also assumed to be caused in the population by “background” processes. The common idea that there is a “radiation cancer” that a physician can diagnose as distinct from any other cancer is false. The assignment of causation—that a cancer was caused by radiation—depends not on a physician's observations but on knowledge of likely exposure. Because the cancers produced by radiation and those produced by background are indistinguishable, it is likely that, at some stage in the tumor formation process, the background and radiation doses act in the same way. This is an assumption that has not been refuted, although of course it is an assumption, the validity of which must constantly be questioned. Whatever the basic biological process relating a dose to cancer, Taylor's theorem then tells us that a differential linearity results, provided that the radia-

tion dose acts similarly to background. This argument is a general one and can apply to outcomes other than cancer. It can also apply to other cases such as respiratory problems caused by air pollution or cigarette smoking. This then makes it mandatory for any discussion of behavior related to low doses to include a discussion of what causes the natural background of cancers. It should also have implications for public policy and public perception. It is usually assumed that chronic radiation sickness and the effect on growth and development (Section 3.10) are threshold-type effects following the dose-response curve of Fig. 1. But the general argument applies to more than cancer. Either of these two medical outcomes could follow a linear dose response. In the Techa River study of chronic radiation sickness, for example, high radiation exposure was explicitly included in the definition of the disease, making it impossible to study the dose-response below that screening level.

The cancers from introduced radiation and background radiation cannot be distinguished by pathologists with typical, currently available tools. But it is possible that they can be distinguished, in principle, in which case the preceding argument would not apply. One, but only one, study suggests that the chromosome structure of an acute myelogenous leukemia (AML) occurring subsequent to, and presumably caused by, radiotherapy was appreciably different from cases of AML that occur “naturally.” If this turns out to be a general result, the low-dose extrapolation arguments must be drastically reconsidered.

5.6 Epidemiological Studies of Low-Dose Behavior

There is great difficulty in measuring any effect, positive or negative, at the low radiation doses at issue. First, a study of 10,000 patients is only just enough to measure an effect, either beneficial or adverse, of 6% with a statistical significance of three standard deviations. Second, at levels of a few percent, other nonstatistical errors enter into any study. To measure reliably an increase in cancer rate of 10% would necessitate careful study of nearly 1000 patients with the disease. Another important effect is that if there is differential uncertainty in measuring the radiation dose, any structure in the dose-response curve tends to get washed out, and a threshold-type curve may appear to be linear.

The earliest studies of the survivors of the atomic bombs in Hiroshima and Nagasaki suggested (with

weak statistical significance, because they are on the edge of the natural background) a threshold at a dose of about 0.2 Sv (= 20 rem), or even hormesis. The data from the most recent follow up are better fit with linearity down to a dose of 0.05 Sv (= 5 rem). The recent analysis takes all the data together in a manner similar to a maximum-likelihood analysis. It is well known that a maximum-likelihood approach can pull out of the “noise” those signals that are not clear when binned data are taken, and a Gaussian approximation is used. It is, however, admitted that there may well be nonstatistical errors that invalidate this conclusion. The authors of the recent research warn against concluding that there is any health effect of radiation, adverse or beneficial, below 0.1 Sv (< 10 rem).

One of the first studies to suggest a threshold involved the study of bone cancers among the radium dial painters of the early 20th century. The dose to which they were exposed can be measured very reliably after diagnosis, either *in vivo* or in autopsy, because the radium stays in the bone. Two groups of dial painters have been studied, European and American. Data from the U.S. study (Fig. 3) indicated no cancers at a low dose. Linearity is statistically excluded. However, this is for alpha particle exposure and may not be generally applicable.

Workers at nuclear power plants and workers at other fuel cycle facilities have also been exposed to radiation at levels above background. A United Nations agency, the International Agency for Research on Cancer (IARC), compiled a database of 90,000 workers in the nuclear industries of the United States, Canada, and the United Kingdom. In their compilation, several workers received cumulative doses up to 0.2 Sv (= 20 rem). The data alone do not establish a clear link to any form of cancer. This is not merely a “meta-analysis” of several papers but is a combined study of those groups for which the data are deemed reliable. The additional number of leukemias seen is consistent with an extrapolation from the number at higher doses (but the lower 95th percentile of the number is close to zero). The number of additional “solid” cancers is close to zero but the upper 95th percentile is close to the linear extrapolation from higher doses. This suggests that numbers five times larger, as sometimes suggested by individual studies at a single location, are definitively excluded.

Data on workers who helped to clean up Chernobyl and live in the Russian Republic have been collected by the Medical Research Laboratory in Obninsk, Russia. There is a small, marginally

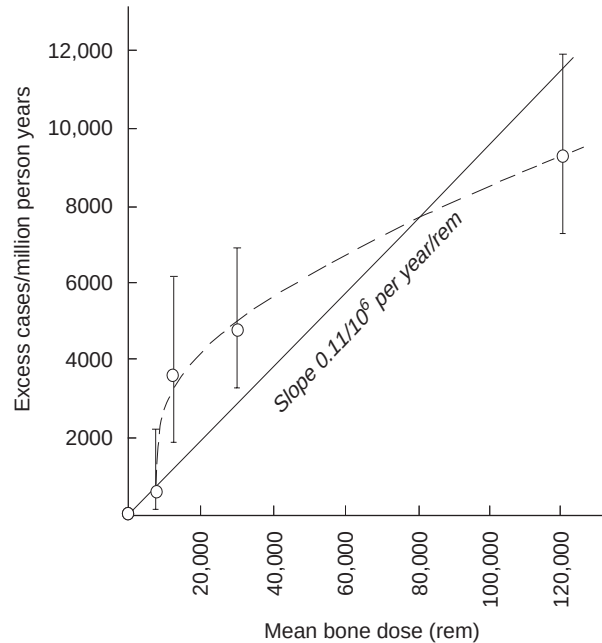


FIGURE 3 Dose-response data for a bone cancer study at Argonne National Laboratory. Study participants were exposed to radium-226 during a 20-year period (1915–1935); follow-up assessments were continued through 1971. The group included watch dial painters and some patients given radium therapeutically. Error bars include the 90% range.

significant, increase in leukemia, in agreement with the calculated linear response. Data on lung cancer among another group of cohorts, the previously discussed Mayak workers, are particularly interesting. The first studies on the Mayak workers suggested that the dose-response relationship was quadratic with dose (in qualitative agreement with animal data), and therefore a low-dose effect approaching zero seemed to make sense. But a more careful look at the data suggests that a linear dose-response relationship fits the data better. However, the doses are not low and a threshold or reduced effect at lower doses is possible.

Attempts have been made to use data on therapeutic uses of radiation to address the issue. There are few such studies with a large enough data sample. Some of these suggest that there is a dose below which there is no effect. But there are potential biases in such a study in that physicians may only look at cases in which the doses are high. There is no consensus that systematic errors are so well controlled that the studies can be used to discuss directly the shape of the dose-response curve below a total dose of 0.2 Sv (= 20 rem).

Only recently have there been studies of the effects of radon on people in residential situations. There are

two types of studies. One, an “ecological” study, compares the average lung cancer rate in a community to the average radon concentration in the houses of that community. The average lung cancer rate decreases with increasing radon concentration. It would be a logical non sequitur to derive directly from such a study the relationship of the probability of an individual person succumbing to cancer and the radon concentration to which that individual is exposed (the dose–response relationship). To make such a conclusion is sometimes called “the ecological fallacy.” However, it is legitimate to compare any set of data with a theory, and if the data do not fit, the theory must be wrong. In particular, the theory espoused by the U.S. EPA, which uses a linear dose–response relationship, cannot be correct. However, the EPA error may be in not including other risk factors for lung cancer that are more important than radon. Another type of study is the “retrospective cohort” approach. There are a few such studies, in which a group of people are followed and the individual doses from radon are estimated. These are free from the ecological fallacy, but there are few data in the low-dose region, to which 90% of Americans are exposed. It is important to realize that any conclusion about the risk at low doses (that is, doses below natural background level) derived from these studies is dependent on an extrapolation, which may not be in direct disagreement with the ecological studies that use more people and are therefore statistically more significant. Most experts agree that below a radon concentration of about 20 pCi/liter, no one is sure which type of study is correct.

6. WEIGHTING FACTORS (W_r), OR RELATIVE BIOLOGICAL EFFECTIVENESS

As a first approximation, the effect of radiation on people depends only on the energy deposited per unit volume (measured in grays or rads). However, radiobiological reasoning suggests that particles traveling at speeds less than the velocity of light are relatively more effective at producing radiation effects, and a weighting factor, W_r , sometimes called the radiobiological equivalent (RBE), must be used. For high-energy X ray and gamma ray exposures, $W_r = \text{RBE} = 1$. Most of the studies of effects of radiation on people are associated with ionization by particles traveling at speeds near the velocity of light, and the ionization per unit of energy loss is at a minimum. For example, the radiation from the

TABLE I

Approximate Radio Biological Equivalent (RBE) Factors (Simplified)

Radiation type	RBE factor
Photons	
E > 250 keV	1.0 (definition)
30 keV < E < 250 keV	1.9
E < 30 keV	2.4
Electrons	
E > 60 keV	1.0
15 keV < E < 60 keV	1.9
E < 15 keV	2.7
Neutrons	
E > 20 MeV	2.2
2 MeV < E < 20 MeV	4.4
0.1 MeV < E < 2 MeV	7.8
10 keV < E < 100 keV	3.4
E < 3.4 keV	1.7

Modified from the SENES study (see Further Reading) for probability of causation.

Hiroshima and Nagasaki bombs consisted mostly of high-energy gamma rays. Data on the correct value to use for this factor for other situations are sparse. The factors are usually estimated from animal data and contain large errors. This is clear from Table I, which summarizes the latest data.

7. POLICY IMPLICATIONS

7.1 ALARA

The first step in discussing a sensible public policy is to summarize the data in quantitative terms. This is repeatedly done by various agencies. In the United States, the important agencies include the National Council on Radiation Protection and Measurements and the National Academy of Sciences, Committee on Biological Effects of Ionizing Radiation. In the United Kingdom, the agencies include the National Radiological Protection Board. Internationally, there is the United Nations Subcommittee for the Effects of Atomic Radiation. These organizations produce a set of predictions that, for want of a better word, is often called the “establishment” position. This establishment position assumes, for example, that the dose–response curve is linear at low doses, and provides data on the best estimate of the slope of this dose–response curve. As noted previously, this assumption remains unproved and is probably unprovable in any direct way.

The details of the establishment predictions are complex. The predictions depend on location of exposure, age at exposure, duration of exposure, rate of exposure, and gender. They concentrate almost exclusively on effects on cancer, but there are differences between the various cancer body sites. Scientists, however, tend to summarize the effects by a risk coefficient that is a sum over all fatal cancers, a weighted average over the population and for exposure to whole body irradiation. The cancer risk then is about 1/20 to 1/50 of lifetime cases per person-sievert (1/2,000 to 1/5,000 per person-rem). Alternatively stated, the dose to add 1% to the cancer incidence (which is a relative risk of 0.03) is between 0.2 and 0.5 Sv (20–50 rem). Of course, for some cancer body sites and for some ages the relative risk (RR) is much greater.

Starting in 1927, the International Commission on Radiological Protection (ICRP) noted that no threshold had been found below which adverse effects on health are known not to occur. They recommended that “no radiation exposure shall be permitted without expectation of some benefit,” a view that survives today. Attempts have been made over the years to define the ICRP precautionary principle more clearly. It was first suggested that radiation exposures be kept as low as practicable (ALAP). In the early 1970s, scientists, unhappy with the vagueness of the terms, replaced ALAP with “as low as reasonably achievable” (ALARA) (economic and other factors taken into account). Europeans instead adopted “as low as reasonably practicable” (ALARP). It is unclear why anyone thought that this wording was an improvement.

In 1975, after a 2-year-long public hearing, the Atomic Energy Commission (later the Nuclear Regulatory Commission) attempted a quantitative definition of ALARA. It decided that radiation dose should be reduced if the cost per unit of integrated radiation dose (integrated over the whole population) is less than \$1,000 per person-rem. (The present limit, which has been updated in response to political correctness, a change of units of radiation exposure, and inflation, is \$200,000 per person-sievert.) Using a widely used establishment slope of the dose-response relationship for ionizing radiation and cancer, this corresponds to a cost of about \$6,000,000 per cancer averted. It is interesting to note that the U.S. EPA has recently, and apparently independently, come up with a number of \$6,100,000 as the “value of a statistical life” to be used in discussing costs to avert a cancer, for which the EPA also assumes a linear low-dose response.

On face value, this rule would suggest that an attempt to reduce exposure that costs more than \$200,000 per person-sievert is overreacting to the radiation risk, and also that any such attempt should be enforced only “on the average.” The reality is more complicated. In disposal of nuclear wastes, for example, this rule has not been applied and already much more money has been spent than the rule suggests. Also, it is clear that exposure to radiation of someone aged 70+ years is far less important than it is to someone aged 40 years. A 70-year-old has a shorter time to live, and the 20-year latent period for most cancers almost guarantees that he or she will not get a radiation-induced cancer.

Moreover, people have traditionally been allowed higher exposures in their work, with the obvious compensation of having a job. Thus, the National Council for Radiation Protection and Measurement (NCRPM) recommended in 1990 that the rule for occupational exposure be closer to \$2,000 per person-sievert (\$20 per person-rem). The radiation protection authorities are just beginning to consider such niceties, which are easily handled by standard tools of decision analysis. Despite these complexities, the application of “precautionary” policies for radiation protection has been closely tied to cost-benefit analysis, and to the conservative (but difficult-to-test) assumption that the risks from exposure at low levels is linearly proportional to the exposure.

Professional societies, unhappy with the present use of ALARA for very low doses, have emphasized the uncertainty surrounding the possible effects caused by low doses of radiation. For example, the Health Physics Society issued the following statement:

In accordance with current knowledge of radiation health risks, the Health Physics Society recommends against quantitative estimation of health risks below an individual dose of 5 rem (= 0.05 Sv) in one year or a lifetime dose of 10 rem (= 0.1 Sv) in addition to background radiation. Risk estimation in this dose range should be strictly qualitative, accentuating a range of hypothetical health outcomes with an emphasis on the likely possibility of zero adverse health effects. The current philosophy of radiation protection is based on the assumption that any radiation dose, no matter how small, may result in human health effects, such as cancer and hereditary genetic damage. There is substantial and convincing scientific evidence for health risks at high dose. Below 10 rem (which includes occupational and environmental exposures), risks of health effects are either too small to be observed or are nonexistent.

7.2 Probability of Causation

When a person has cancer, society or the individual wants to know the cause and who to blame. In most

cases, the cause is, of course, not known and there may be multiple causes. It is not appropriate to calculate the “risk” of getting cancer (or indeed any other disease) when it is already known that the person has cancer, because the risk in such a situation is unity! But it is appropriate to use a different concept, the “probability of causation.” Loosely, the question to ask concerns what the probability is that the cancer was caused by the particular agent of concern. The answer can be derived from the general ideas of probability using the concepts of conditional probability. The probability of causation (POC, or sometimes PC) is defined as follows:

$$\text{POC} = \frac{\left(\begin{array}{c} \text{Risk of cancer from the} \\ \text{cause under study} \end{array} \right)}{\left(\begin{array}{c} \text{Risk of cancer from all causes} \\ \text{including the cause under study} \end{array} \right)}$$

There have been many theoretical criticisms of this concept and some scientists prefer the expression “assigned share.” If the radiation effect is beneficial, the probability can be negative, which makes no sense. If radiation is only one of several causes, the sum of the various POCs for different possible causes can be greater than unity. But with the approximations usually used, POC makes sense. Synergistic effects are usually ignored, and the dose-response slope is linear and positive. Then the probability is positive. To evaluate this, the risk from causes other than the one under consideration is usually taken as the natural incidence of the cancer in question the assumption is then made that the population group from which the incidence data are derived is representative of the individual whose risk is evaluated in the numerator. Dividing through by the natural incidence, we derive the equation $\text{POC} = \text{ERR}/(1 + \text{ERR})$, where ERR is the excess relative risk for the cancer in question. Of course, ERR is a function of age and gender.

In the 1950s, the United States conducted many atomic bomb tests in the atmosphere in Nevada. In some tests, U.S. army veterans were deliberately situated as observers to give them experience and ability to respond in battle without panic. In 1985, some of these veterans, and, later, residents of the states downwind of the test explosions, put pressure on Congress to give them compensation for the assumed damage. Because, as noted previously, there is no such thing as a “radiation cancer,” the U.S. Congress requested the National Cancer Institute to develop a set of tables to determine the probability that a person’s cancer was due to radiation exposure.

These tables were published, but have since been updated into a user-operated computer program. The tables assume a linear dose-response relationship. There is disagreement on how to use the POC concept. In the United Kingdom, partial compensation is given to persons whose POC is greater than 20%. Procedures in the United States are governed by the legal maxim that “it is more likely than not” that a particular action or dose was the cause of the injury. This translates to $\text{POC} > 50\%$. But few veterans would be compensated under such a rule, and it was desired politically to be generous to the veterans. An uncertainty distribution is now requested and the screening demands that only if POC is $< 50\%$ with 99% certainty shall the effect be ignored.

A more logical approach would be to note that for an individual it does not matter whether a cancer developed because of bad luck in the stochastic probability or because the actual risk coefficient was on the high side of the uncertainty distribution. Then the two distributions should be combined (folded) into one. This has not been used in any situation to date.

8. PUBLIC PERCEPTION

As already noted, public perception of the effects of radiation exposure changed dramatically in the second half of the 20th century, presumably as a result of the atomic bombs in August 1945. This, combined with an unattainable desire for “zero risk,” leads to considerable logical problems. For example, it has been proposed in one jurisdiction that in dismantling any building containing radioactive material, nothing with any measured radiation levels shall be placed in an ordinary landfill. If that were applied to cemeteries, it would be impossible to bury anyone, because radioactivity in a body can be easily measured. Nor could anyone be cremated, because radioactivity in the effluent would cause the crematorium to be shut down. However, zero risk is usually used to apply only to a particular action or process a particular person or group dislikes, for which a different logic applies.

SEE ALSO THE FOLLOWING ARTICLES

Air Pollution from Energy Production and Use • Air Pollution, Health Effects of • Electromagnetic Fields,

Health Impacts of • Gasoline Additives and Public Health • Nuclear Proliferation and Diversion • Nuclear Waste • Occupational Health Risks in Nuclear Power

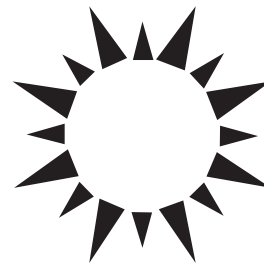
Further Reading

- Boice, J. D., Jr., and Fraumeni, J. F., Jr. (1984). "Radiation Carcinogenesis: Epidemiology and Biological Significance." Raven Press, New York.
- International Atomic Energy Agency (IAEA). (1997). "Doses of Ionizing Radiation: Biological Effects and Regulatory Control." IAEA-TECDOC-976, IAEA-CN-67/63, pp. 223–226. IAEA, Vienna.
- International Atomic Energy Agency, Vienna, Austria. Web site: <http://www.iaea.org> (Accessed in March 2003.)
- International Committee on Radiation Units and Measurement. Web site: <http://www.icru.org> (Accessed in March 2003.)
- Mettler, F. A., and Upton, A. C. (eds.). (1995). "Medical Effects of Ionizing Radiation," 2nd ed. W. B. Saunders, Philadelphia.
- National Council on Radiation Protection and Measurements (NCRPM). (1995). "Principles and Application of Collective Dose in Radiation Protection." Report No. 121. NCRPM, Bethesda.
- National Council on Radiological Protection and Measurements, United States. Web site: <http://www.ncrp.com> (Accessed in March 2003.)
- National Radiological Protection Board, United Kingdom. Web site: <http://www.nrpb.org.uk> (Accessed in March 2003.)
- National Research Council, U.S. National Academy of Sciences. (1990). "Health Effects of Exposure to Low Levels of Ionizing Radiation." BEIR V, 1990. National Academy Press, Washington, D.C.
- Pierce, D. A., Shimizu, Y., Preston, D. L., Vaeth, M., and Mabuchi, K. (1996). Studies of the mortality of atomic bomb survivors. Report 12, Part I. Cancer: 1950–1990. *Radiat. Res.* **146**, 1–27. Radiation Effects Research Foundation, Japan. Web site: <http://www.rerf.or.jp> (Accessed in March 2003.)
- SENES, Oak Ridge, Inc., Center for Risk Analysis. Probability of causation calculation and numbers. Web site: <http://www.senes.com> (Accessed in March 2003.)
- Shapiro, J. (1990). "Radiation Protection: A Guide for Scientists and Physicians." 3rd ed. Harvard University Press, Cambridge.
- United Nations Scientific Committee on the Effects of Atomic Radiation. (2001). "Report to the General Assembly, with Scientific Annex." UNSCEAR 2001. United Nations, New York.

Rebound Effect of Energy Conservation

HORACE HERRING

The Open University
Milton Keynes, United Kingdom



1. Implications of the Rebound Effect
2. Price Elasticity Effects
3. Transformation Effects
4. Theoretical Argument
5. Empirical Evidence
6. Policy Implications
7. Conclusion

Glossary

energy services The services that consumers want from the supply of electricity and other fuels: warm homes, cold beers, lighting, mobility, etc. A combination of energy supply and appliances (boilers, refrigerators, lamps, cars, etc.) provide these services.

price elasticity The ratio of the percentage change in energy use to the percentage change in energy price.

The rebound effect (or take-back effect) is the term used to describe the effect that the lower costs of energy services, due to increased energy efficiency, have on consumer behavior both individually and nationally. Simply stated, the rebound effect is the extent of the energy saving produced by an efficiency investment that is taken back by consumers in the form of higher consumption, in the form of either more hours of use or a higher quality of energy service.

1. IMPLICATIONS OF THE REBOUND EFFECT

Implications of the rebound effect are demonstrated by the following example: If a 18-W compact fluorescent bulb is replaced by a 75-W incandescent

bulb, the energy saving should be 76%. However, this is seldom the case. Consumers who realize that the light now costs less per hour to run are often less concerned about switching it off; indeed, they may intentionally leave it on all night. Thus, they “take back” some of the energy savings in the form of higher levels of energy service (more hours of light). This is particularly the case when the past level of energy services, such as heating or cooling, was considered inadequate.

The rebound effect is a phenomenon based on economic theory and long-term historical studies, but as with all economic observations its magnitude is a matter of considerable dispute. Its significance for energy policy has increased during the past few decades with the claim by energy analysts in the 1970s, and later by environmentalists in the late 1980s, that increasing energy efficiency would lead to reduced national energy consumption and hence lower greenhouse gas emissions. Whether this claim is feasible depends crucially on the extent of the rebound effect: If it is small ($<100\%$) energy efficiency improvements will lead to lower energy consumption, whereas if it is large ($>100\%$) energy consumption will be higher. Note the use of the relative terms lower and higher: What exactly they are relative to has often been left unstated and has been a cause of much confusion in energy policy debates. Sometimes it refers to current energy consumption, whereas at other times it refers to a reduction in the future rate of growth in energy consumption.

The claim that increasing energy efficiency would lead to reduced national energy consumption was first challenged by Len Brookes in 1979 in his review of Leach's pioneering work, “A Low Energy Strategy for the UK,” when he criticized Leach's approach to estimating national energy savings because of its

failure to consider macroeconomic factors. This was followed in the early 1980s by similar criticism by Daniel Khazzoom of the work of Amory Lovins. The criticism of Brookes and Khazzoom was called the Khazzoom–Brookes (KB) postulate by the economist Harry Saunders in 1992. The KB postulate (sometimes referred to as the Jevons paradox) may be described as follows: Those energy-efficiency improvements that, on the broadest considerations, are economically justified at the microlevel lead to higher levels of energy consumption at the macrolevel than in the absence of such improvements.

The debate grew more intense in the early 1990s, spurred by global warming concerns and policy debate over the role of energy efficiency. It was conducted mostly in the pages of the academic journals *Energy Policy* and *Energy Journal* but also spread to the pages of *The New York Times* in late 1994 and the leading UK science magazine *New Scientist* in 1998. It culminated in two special issues of journals in 2000 devoted to aspects of the rebound effect—*Energy Policy* and *Energy and Environment*—to which many of the protagonists in this debate contributed. There is no dispute that the rebound effect exists; what is at the core of the conflict is its magnitude, and the two opposing positions can be summarized as follows:

1. Energy use is higher than if there had been no efficiency response—a position maintained by Len Brookes and, in some circumstances, Harry Saunders.
2. Energy use is lower than if there had been no efficiency response—a position maintained by Lee Schipper and colleagues.

Each side has supported its case with a mix of theoretical argument and empirical observations, based on historical data on energy use. A key problem in resolving the two positions is that it is not possible to perform control experiments to determine whether energy use is higher or lower than if there had been no efficiency improvements; after all, there is only one past. Attempts to estimate the overall magnitude of the rebound effect, using theoretical economic models based on neoclassical growth theory, have proved inconclusive, with the results dependent on assumptions about the elasticity of substitution of energy for other factors of production—the greater the elasticity, the greater the rebound. Another problem is that the rebound effect differs by energy end use and sector of the economy, and the response at the microeconomic level (the consumer) is different from that at the macroeconomic level (the national economy). Also,

the rebound effect can vary between countries, depending on the energy costs and unmet demand for energy services; thus, it may be considerably higher in developing countries than in the United States.

However, all agree that the rebound effect is closely linked to the elasticity of substitution, which is a component of the widely measurable phenomenon of price elasticity. The higher the observed price elasticity of energy services, the greater the rebound effect. This leads to the paradoxical position that any imposition of energy or carbon taxes, in the hope of reducing energy use, would have its impact severely blunted by the rebound effect if consumers have a high price elasticity.

2. PRICE ELASTICITY EFFECTS

Rebound effects are due to the increased use of energy services caused by the reduction in their effective price due to greater efficiency (price elasticity) and can be classified into five categories:

1. Direct effects due to the desire of consumers to use more of any commodity due to its lower price
2. Income-related effects due to the fact that with lower energy price, more income is available to spend on energy
3. Product substitution effects due to the inclination to substitute energy services for other final consumption goods when energy services become less expensive than other final goods
4. Factor substitution effects due to the inclination to substitute energy for other factors of production (capital, labor, materials, time, etc.) in the production of final goods
5. Transformational effects due to long-term changes in the economy caused by changes in technology, consumer preferences, and even social institutions brought about by the substitution of energy for other factors of production.

The effect of perceived lower costs on energy use is termed price elasticity, which is the ratio of the percentage change in energy use to the percentage change in energy price. Price elasticities vary by commodity and over time, depending on the ability of consumers to respond to price changes through changes in behavior, substitution of alternatives, or technical change. It is important to distinguish price-induced efficiency gains from non-priced-induced gains. The former are caused by factor substitution and may involve costs to the economy (such as was caused by the oil price increases in the 1970s and

may result from the imposition of energy taxes). Non-priced-induced gains are due to technological improvements, and the rebound effects from them are of most concern to energy analysts.

Much evidence on the magnitude of the direct rebound effect derives from U.S. transportation studies for which there is good statistical data on miles traveled per vehicle and petrol consumption. The results indicate that the number of vehicle miles traveled will increase (or rebound) 10–30% as a result of improvement in fuel efficiency. Similar results are obtained for domestic energy services, such as space heating (Table I). However, the long-term price elasticity is considerably higher because the ability of consumers to respond to changes in prices is much greater: For private transport in Europe, the rebound (or long-term price elasticity) is estimated to be 30–50%.

Expansion of service demand can also take the form of consumer preference for higher levels of quality, comfort, and performance. For instance, the oil price increases of the 1970s stimulated R&D on more efficient car engines providing more miles per gallon of gas (petrol); when oil prices declined in the late 1980s car manufacturers offered consumers more powerful and better equipped cars (e.g., sport utility vehicles) with the same fuel costs but lower miles per gallon. Here, consumers appear to take the possible savings in the form of higher levels of performance or quality. Another area in which consumers “trade up” is substituting energy for time: faster cars and planes, labor-saving domestic appliances, fast food, and private rather than public transport.

An important impact of the effect of higher efficiency, and consequent lower cost of energy services, is on marginal consumers—those who could

not previously afford the service. The appliance market, such as for air-conditioning, will expand as the cost of energy service declines, with the magnitude depending on the elasticity of service demand. This is because consumers’ choice of the efficiency of an equipment depends on factor costs, and more efficient equipment will increase the demand for the service because of lower marginal costs per unit of service.

The second effect on consumers of lower energy costs is the income effect—their ability to spend the monetary savings on other goods and services in the economy, which will contribute to economic growth. What they spend this discretionary income on depends on their current income levels: Those with low incomes will use it for basic goods and those with higher incomes on luxury services. However, almost all goods and services involve some energy consumption in their production, distribution, and consumption, and their energy intensity varies markedly. Schipper argues that the income effect is very low because only 5–15% of income saved is spent on indirect energy use (with air travel at the high end).

What the monetary savings is spent on is crucial to the concept of sustainable consumption—that is, consumption that has low environmental impact (i.e., low material and energy content).

3. TRANSFORMATION EFFECTS

Probably the greatest effect (in the long term) of lower costs of energy services is on the direction and pace of technical change and innovation in the economy. New goods and services will be introduced to take account of the possibility of lower costs, and mass markets will be created by the continuing decline in operating costs bought about by efficiency improvements. As Len Brookes comments.

It is inconceivable that we should have had the high levels of economic output triggered by the industrial revolution if energy conversion had stayed where it was at the beginning of the 19th century. Energy productivity and the productivity of other factors of production fed on one another with rising energy efficiency contributing to rising productivity of other factors of production—labor and capital—and rising output contributing to rising energy efficiency by way of embodied technical progress. Without this interactive process we should not have had, in the meantime, the very large increases in energy consumption alongside large improvement in energy conversion efficiency.

An illustration of these very large increases in energy consumption and energy efficiency due to

TABLE I

Estimates of the Direct Rebound Effect by End Use for the U.S. Residential Sector^a

End use	Rebound effect (%)	No. of studies
Space heating	10–30	26
Space cooling	0–50	9
Water heating	<10–40	5
Lighting	5–12	4
Car transport	10–30	22

^aFrom Greening, Greene, D., and Difiglio, C. (2000). Energy efficiency and consumption—The rebound effect—A survey. *Energy Policy* 28, 398.

technical change is given here for public lighting in Great Britain for the period 1920–1995. This sector is chosen because there are good statistical series for electricity consumption, lamp efficiency, and road mileage. These statistics show an increase of more than 30-fold in electricity consumption for public lighting and 20-fold in lamp efficiency but less than 50% in road mileage. The net result of this increase in consumption and efficiency since the early 1920s is that light intensity, in terms of lumens per mile, increased more than 400 times and increased 4-fold since 1960. During the same period, energy intensity, in terms of MWh per mile, increased 25-fold, with a 250% increase since 1960 (Fig. 1).

Thus, it seems that most of the tremendous increase in lamp efficiency has been in the form of higher levels of service, both in more miles illuminated and in higher illumination levels, and not in the form of lower consumption. Interestingly, both consumption and efficiency seem to have leveled off in the 1990s, whereas during a period of rapid growth in consumption from 1960 to 1980, lamp efficiency increased by 50%.

This process of technical change is aided by manufacturers that continually seek efficiency improvements in order to lower consumer costs and create new mass markets. In conjunction with this, utilities welcome efficiency improvements to lower the cost of their product: Better to sell much at a low profit margin than little at a high margin. The massive expansion of electricity consumption in the 20th century was fueled by the continual decrease in electricity prices, bought about by a nearly 10-fold

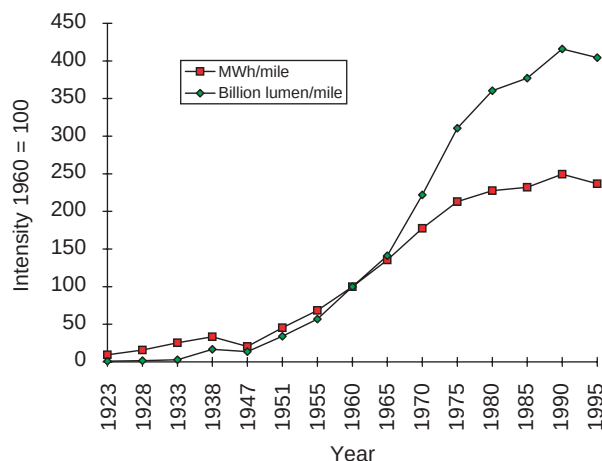


FIGURE 1 Changes in energy and light intensity for public lighting in Great Britain, 1920–1995. Reprinted from Herring, H. (1999). Does energy efficiency save energy? The debate and its consequences. *Appl. Energy* 63, 209–226, Copyright 1999, with permission from Elsevier.

increase in generation efficiency, which spurred the development of new electrical goods and services: electric lighting in the 1900s, domestic refrigeration in the 1930s, TV in the 1950s, microwaves and videos in the 1980s, and computers and the Internet in the 1990s.

How much the improvement in energy efficiency per se stimulates economic growth is the subject of much discussion, and it is explored later.

4. THEORETICAL ARGUMENT

Key contributors in the debate over the rebound effect are economists, who seek to identify through economic theory and mathematical models the conditions under which the rebound effect may be greater than 100% (a state termed backfire). Debate by economists over the effect of efficiency improvements is long-standing, and current economists all acknowledge the work of one of the founding fathers of energy economics, Stanley Jevons, who in his classic work “The Coal Question,” first published in Great Britain in 1865, argued that “it is a confusion of ideas to suppose that the economical use of fuel is equivalent to diminished consumption. The very contrary is the truth.”

Jevons pointed out that

the reduction of the consumption of coal, per ton of iron, to less than one-third of its former amount, was followed, in Scotland, by a 10-fold increase in total consumption, between the years 1830 and 1863, not to speak of the indirect effect of cheap iron in accelerating other coal-consuming branches of industry.

Thus, Jevons laid the foundation for the idea that the rebound effect (in energy markets) could be greater than 100%, which is sometimes referred to as the Jevons paradox. The conditions under which it may occur are explored by the Austrian economist Franz Wirl and the U.S. economists Harry Saunders and Richard Howarth using extensive mathematical notion. Wirl, in his book “The Economics of Conservation Programs,” argues that the extent to which more efficient appliance raises the service demand depends on elasticity of service demand:

The most important consequent is that (ceteris paribus) a more efficient appliance raises service demand such that actual conservation falls short of an engineering efficiency comparison. Indeed “conservation” in the sense of improving energy efficiency may raise energy demand, if demand for service is elastic.

He demonstrates that whereas higher prices reduce unambiguously demand, the impact of efficiency is

ambiguous and an improvement in efficiency need not lower consumption. He argues, "A necessary and sufficient condition for excluding this so-called conservation paradox is that service demand is inelastic, which in turn is equivalent to the requirement that the elasticity of marginal benefit is less than -1 ."

He concludes that the exclusion of his conservation paradox requires a price-inelastic service demand, as for price-elastic services an efficiency improvement of 1% will lower the cost of the service by 1%, which will induce an expansion of the service of more than 1%. Hence, the expansion of service more than outweighs the conservation due to the efficiency improvement. Thus, for the rebound effect to be less than 100%, Wirl concludes that the elasticity of marginal benefit has to be less than -1 (i.e., elasticity of service demand is less than 1, which is inelastic).

Wirl observes that the rebound from efficiency improvement on electricity demand works through increasing the service demand. However, in practice some of this feedback may not lead to a physical expansion of the service demand but to upgrading, for example, cars with more features and power (larger and faster). He makes a crude assessment of this rebound effect, arguing that it is small for energy services close to saturation levels (and with low marginal costs), such as stereos or TVs, but larger for services with substantial marginal costs, such as space heating, air-conditioning, transport, and perhaps lighting. However, in the former case the absolute conservation potential is small, whereas in the latter the potential is large. This reflects the fact that consumers are more sensitive (and responsive) to end uses having a large energy cost. Thus, Wirl concludes that the rebound effect seems important for services with a significant conservation potential but negligible for services with a minor conservation potential in terms of kilowatt-hours.

Work by Saunders and Howarth, using neoclassical growth theory, seeks to identify the conditions under which backfire (rebound effect $>100\%$) occurs. The results are sensitive to the choice of production function, and for some functions they depend on the assumptions made about the elasticity of substitution between energy and other factors. The use of the Cobb–Douglas production function, with the assumption of an elasticity of substitution of 1, validates the KB postulate, although many energy analysts question its conceptual underpinnings, which rest on neoclassical economic principles. In contrast, Richard Howarth uses a Leontieff production function that assumes zero elasticity of substitution, and

he argues, based on this model, that improved energy efficiency cannot give rise to increased energy use except under implausible assumptions.

Saunders lists the cases and circumstances in which the KB postulate is or is not validated. He is developing a simple economic model (or first-order tool) that will allow economists to take an arbitrary specified new efficiency technology and deploy it against an arbitrary specified production function to generate a prediction of the resulting change in fuel use, including possible rebound effects. Preliminary results show that energy efficiency technologies that enhance other factors of production may produce backfire, and that there is a marked difference in rebound between industrial sectors and between the same industries in different countries.

The impact of energy efficiency technologies in stimulating the use of other factors of production has been termed the Cashmir effect, with the observation that a technical fix to improve energy productivity can produce a bonus in materials productivity. As Ronald Sutherland comments, "Energy efficiency increases, not as a direct effort to reduce energy use, but as a result of overall productivity improvements in all inputs. In economic terms capital and energy appear to be complements not substitutes."

To resolve this issue is no easy task. Richard Howarth remarks,

Sorting out the empirical dimensions of the Khazzoom–Brookes hypothesis ... would require detailed models that merge engineering approaches to energy efficiency, micro-engineering studies of the demand for energy services, and macroeconomic models of savings and investments. The construction of such models is an ambitious task.

5. EMPIRICAL EVIDENCE

A number of studies have been performed to determine the impact of improved efficiency on energy use. In the United States, a major study on energy processes during the period 1880–1970 was done by Sam Schurr, whereas Lee Schipper and colleagues have done much pioneering work on energy trends for Organization for Economic Cooperation and Development countries on a sectorial level for the period 1970–1995.

However, the problem is in measuring energy efficiency because its two indicators, energy intensity (energy use per unit output) and the energy coefficient (the output elasticity of energy consumption), can give false signals. Schurr maintained that allying capital and labor inputs with new injections

of energy into economic systems can increase the productivity of both capital and labor. This results in a decrease in energy intensity due to a larger denominator in the form of higher economic output. This can deliver a false message when in fact there has been no change in the efficiency of conversion of fuel to useful heat and work. Nevertheless, it is accepted that there is a steady long-term trend in efficiency improvement in the economy due to the vintage effect. That is, the tendency for new plant and appliances to be more efficient than those they replace. Thus, it is limited by the rate of stock turnover and the rate of additions to stock, generally due to economic growth.

Schurr's empirical findings indicated that for the period 1920–1953, new technologies, often using electricity, not only increased the productivity of labor and capital but also improved energy productivity (i.e., reduced energy intensity). Energy efficiency improved at the same time that energy consumption and economic output increased. However, total output grew at a faster rate than the rate at which energy intensity declined, so total energy consumption increased. It was only in exceptional circumstances, such as the 1979 oil price hike, that energy productivity exceeded multifactor productivity, which actually declined in 1979 due to economic recession.

Other work examining U.S. energy data includes that by Bill Hogan and Dale Jorgenson, who studied time series data on energy intensity for a large number of sectors and found that there is a trend of increasing energy intensity, once price effects are carefully taken out.

The extent to which energy efficiency stimulates economic activity, and hence increased energy use, is contentious, but the effect is probably small. An examination of trends in energy intensity in International Energy Agency countries for the period 1970–1995 by Schipper and Grubb concluded that feedback effects were small in mature sectors of mature economies and only potentially large in a few cases. Their thesis is that the improvement in efficiency per se is only a small part of the reason why total energy use may have increased.

They remark that the scale of the system keeps increasing with population, household formation, and increasing incomes and sectorial output; thus, energy use must increase. However, they conclude that energy efficiency improvements, as measured by lowered energy intensities, almost always led to lower use than otherwise, and such energy efficiency improvements do restrain energy growth.

The assumption that a major part of the rebound effect will automatically manifest itself as increased output or activity levels, and thus be visible through higher energy intensities, may be incorrect. The rebound could, as Wirl points out, take the form of higher quality, performance, and comfort. A key problem in resolving the KB postulate, which is at heart a standard empirical question, is that it is not possible to run historical control experiments on society to determine whether energy use is higher or lower than if there had been no efficiency gains. This means that straightforward observation of visible macroeconomic variables, such as energy intensities, cannot provide much explanation. Instead, it may be more useful to use more sophisticated econometric tools, such as generalized technology production and cost functions, to analyze macro- or sectorial-level statistical data.

6. POLICY IMPLICATIONS

Improving energy efficiency has become, in many countries, a key part of the national strategy to deal with global warming and is founded on the belief that it would lead to reduced national energy consumption and hence lower greenhouse gas emissions. There is a long policy history of seeking national improvements in energy efficiency as a solution to environmental problems, but ultimately efficiency gains have been used to encourage economic growth and progress rather than to reduce consumption. The end result has been that economic growth has outpaced the rate of efficiency increase, and total energy consumption has increased. For instance, in the United Kingdom between 1965 and 1998, efficiency (as expressed by energy intensity, a rough proxy) doubled—a factor two improvement. However, gross domestic product increased by 147%, so total energy consumption increased by 25%. Thus, at current rates of efficiency improvement, it is feasible for there to be a factor four (or fourfold) improvement in the next century. However, as the UK Royal Commission on Environmental Protection comments,

There will continue to be very large gains in energy and resource efficiency but on current trends we find no reason to believe that these improvements can counteract the tendency for energy consumption to grow. Even if energy consumed per unit of output were reduced by three-quarters or factor four, half a century of economic growth at 3% a year (slightly less than the global trend for the past quarter century) would more than quadruple output, leaving overall energy consumption unchanged.

This unequal race between energy efficiency and economic growth is well appreciated by many environmental economists and sociologists concerned with resource consumption. For instance, Mathias Wackernagel and William Rees, in their book "Our Ecological Footprint," explicitly make the point that technological efficiency may actually lead to increased net consumption of resources. They conclude

Ironically then, it is precisely the economic gains from improved technical efficiency that increase the rate of resource throughput. Microeconomic reality demands that these efficiency gains be used to short-term economic advantage. Far from conserving natural capital or decreasing ecological footprints, this leads to higher consumption.

Promoting energy efficiency may not necessarily be the best way to save energy or reduce pollution. It may actually encourage energy consumption by conveying the message that consuming increasing amounts of energy is acceptable as long as energy is consumed by technologies that have been deemed efficient. Mithra Moezzi, a U.S. energy analyst, believes that changes in consumer perception and behavior are required, and that overreliance on technical solutions is mistaken. For instance, she advocates that more emphasis should be given to absolute, rather than relative, energy consumption in energy labels and standards. A larger refrigerator may be more efficient but also consumes more than a smaller one. She remarks,

The standards and guidelines inherent in energy policies reveal moral and technical judgments about what is "good" and what is possible. When evaluating any particular standard, guideline, or other message related to energy conservation or efficiency, it is critical to consider not only "energy saved" in a narrow sense, but what underlying messages are being conveyed and how they might affect the cultural perceptions of efficiency and consumption in short and long terms. By choosing specifications that reflect absolute rather than solely relative consumption, policies may encourage an ideological as well as practical shift in perceptions of energy use.

Society has generally preferred technical or economic solutions, and scientists and engineers have instead emphasized the technical possibilities of a shift to less resource-intensive types of consumption (and also the development of nonfossil energy sources). One solution they advocate is the concept of service efficiency, which may be defined as providing a maximum of useful energy services to consumers using the minimum of materials and energy use. There is an extensive literature on this concept and there have been many attempts to design

new types of products that deliver energy services using innovative combinations of market goods and services and household labor.

Service efficiency in itself is not a panacea for sustainable consumption because the gains are easily offset by an increase in the number and variety of products consumed. Also, as research on the cultural aspects of consumption has shown, it is necessary to understand how and why we consume. This is why innovations that are meant to be efficient and to reduce the need for resources often have the opposite effect. As F.-J. Radermacher, a German sociologist, remarked, :

The trap that we have fallen into again and again over the course of technical progress consists of our always using progress on top of whatever went before (the rebound effect). This effect predicts that market forces and humanity's apparently unlimited capacity for consumption will use new technology to convert more and more resources into more and more activities, functions, services, and products.

It is often argued that information technologies such as computers will reduce material and energy consumption through substituting "virtual" for real experiences and goods. However, Finnish sociologists Heiskanen and Pantzar ask,

Will the information superhighway do away with the urge to travel? Will consumers actually substitute one good for another, or will they want to have it all: the television on, the newspaper on the table, and electronic news pointlessly self-scanning as the consumer of all this information dozes on the couch?

They make the comparison with claims made in the early 1980s about the paperless office, which never happened and, in fact, office paper consumption has increased.

7. CONCLUSION

A key debate about energy efficiency concerns the extent of the rebound: To what extent does energy use increase due to efficiency improvements? Does greater efficiency lead to higher or lower energy use than there would have been without those improvements? This debate, however, is difficult to resolve, as it is with all macroeconomic arguments, since we cannot conduct such (economic) experiments on society. Thus, the rebound argument is contentious and difficult to prove either way. However, it may be possible in the long term, using good empirical research based on sound theoretical tools, to reach strong conclusions about whether it is true (or

untrue) in certain places, at certain times, or under certain conditions and thus determine its overall economy-wide impacts, but there will never be absolute scientific certainty.

Nevertheless, the school of argument that links improvements in energy efficiency with increased energy consumption—the KB postulate—has a strong claim based on the historical record dating back to the industrial revolution and the tremendous economic growth that has followed factor 10 efficiency improvements in the steam engine and other power devices. To what extent energy efficiency technologies stimulated economic growth is uncertain, and economic history, evolutionary economics, and institutional economics could shed more light on this issue than theoretical models based on neoclassical growth theory. The KB postulate may be rooted in the vast structural change in technologies, lifestyles, and social institutions that characterized the past two centuries rather than in a future world determined to reduce, or at least curtail, energy use. We need not repeat past patterns of economic growth and energy use, particularly if we practice sustainable consumption and emphasize quality of output and increased leisure time rather than quantity of output.

The rebound effect is strongly influenced by the difference in substitution potentials (elasticities) among all the factors (energy, labor, capital, and materials) and these change over time. It is very likely that energy efficiency technologies that enhance other nonenergy factors will generate substantial rebound, which will be the case in many industries. It will take some time and much research before it is known definitively whether the KB postulate is valid, and the result is unlikely to be clear-cut. It is likely that the extent of the rebound will vary depending on energy end use, industrial sectors, and even countries. Some undoubtedly will show backfire, some just modest rebound, and some may even show superconservation (i.e., a negative rebound with energy use decreasing more than the efficiency increases).

The task for energy researchers is thus to establish exactly where energy efficiency will be most effective

rather than just promote energy efficiency as the solution to global warming.

SEE ALSO THE FOLLOWING ARTICLES

Consumption, Energy, and the Environment • Cost-Benefit Analysis Applied to Energy • Economics of Energy Efficiency • Energy Efficiency, Taxonomic Overview • Information Technology and Energy Use • Lifestyles and Energy • Obstacles to Energy Efficiency • Potential for Energy Efficiency: Developing Nations

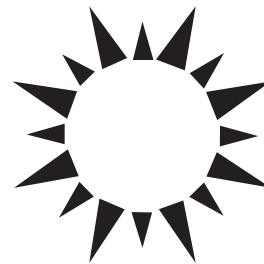
Further Reading

- Binswanger, M. (2001). Technological progress and sustainable development: What about the rebound effect? *Ecol. Econ.* 36, 119–132.
- Brookes, L. (1979). A low energy strategy for the UK by G. Leach *et al.*: A review and reply. *Atom* 269, 3–8.
- Herring, H. (2000a). Is energy efficiency environmentally friendly? *Energy Environ.* 11, 313–326.
- Herring, H. (Ed.) (2000b). How green is energy efficiency? *Energy Environ.* 11(5).
- Howarth, R. (1997). Energy efficiency and economic growth. *Contemporary Econ. Policy* 15, 1–9.
- Khazzoom, J. D. (1980). Economic implications of mandated efficiency standards for household appliances. *Energy J.* 1, 21–39.
- Moezzi, M. (1998). The predicament of efficiency. In “Proceedings of the 1998 ACEEE Summer Study on Energy Efficiency in Buildings, August 1998,” pp. 4273–4282. Available: <http://enduse.lbl.gov/Info/Pred-abstract.html>.
- Sanne, C. (2002). Willing consumers—Or locked-in? Policies for a sustainable consumption. *Ecol. Econ.* 42, 273–287.
- Saunders, H. (1992). The Khazzoom–Brookes postulate and neoclassical growth. *Energy J.* 13, 131–148.
- Schipper, L. (Ed.) (2000). On the rebound: The interaction of energy efficiency, energy use and economic activity. *Energy Policy* 28(6/7).
- Schurr, S. H. (1990). “Electricity in the American Economy: Agent of Technological Change.” Greenwood, Westport, CT.
- Sutherland, R. (1996). The economics of energy conservation policy. *Energy Policy* 24, 361–370.
- Wirl, F. (1997). “The Economics of Conservation Programs.” Kluwer, Dordrecht, The Netherlands.

Recycling of Metals

ERNST WORRELL

Lawrence Berkeley National Laboratory
Berkeley, California, United States



-
1. Introduction
 2. Production of Metals
 3. Recycling of Metals
 4. Energy Use and Recycling
 5. Future of Recycling

Glossary

home scrap Scrap produced at the metal-producing plant as the result of inefficiencies or material losses in the metals-producing plant.

new scrap Scrap that is produced at a manufacturing industry using metal in its production process.

old scrap Scrap recovered from a waste stream after consumption of the metal-containing products. This is also called postconsumer scrap.

primary production The production of metals from primary resources (i.e., ore).

recycling rate The relative amount of metal recycled as a function of apparent domestic consumption of the metal. Apparent consumption is the sum of domestic production and imports minus exports of the metal.

secondary production The production of metals from recovered resources such as scrap.

Metals are used in virtually all applications and are produced in virtually all countries. The production of metals consumes approximately 7% of all the energy in the world. The most important metals are iron and steel, aluminum, copper, and a few others. Metals are produced from primary resources (ore) and from secondary sources such as scrap metal. Because of the high energy intensity and costs of metals, metal recycling is an efficient and effective way to reduce energy consumption in the metal industries. Recycling rates vary for the various metals. Energy savings vary for the various metals depending on the type of metal, amount of metal recycled, and type of scrap recycled.

1. INTRODUCTION

Metals have contributed to the technological progress of society for an extraordinary long time and are still a key material in our society. This is unlikely to change in the future. However, the source of the metals has and will change in the foreseeable future.

Historically, metals had such a high value that they were reused and recycled. However, in modern society increasingly more metals are used in a dissipative way. The use of metals has also rapidly increased throughout society. This had led to the use of lower concentrations of ores for the production of metals. For example, copper is a relatively scarce metal. At the beginning of the 20th century, copper concentrations in ore were more than 2%. Currently, copper concentrations of mined ores are approximately 0.5% or less. This means that more than four times as much ore must be processed to produce the same amount of copper. The use of lower quality resources may have led to increased costs of these materials if it were not for technological progress in mining, ore beneficiation, and metal production technology. However, with the lower concentrations found in ores used today the potential for environmental pollution has increased. It is estimated that annually more than 6.4×10^9 metric tons of ore are extracted from the earth, of which more than 90% is discarded as mine and mill wastes. With the increasing volume of ores used, the amount of energy needed to extract and process the ores has also increased dramatically. This is unsustainable from an environmental and (long-term) economic perspective.

Recycling allows recovering part of the metals entering the industrial society, reducing the need for the extraction of resources from the natural environment. Recycling will save natural resources, reduce pollution, and save energy. The energy savings depend on the energy needed to recover and upgrade the metals from the waste streams compared to the

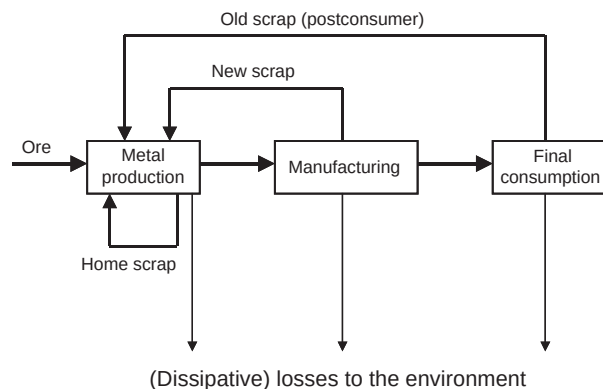


FIGURE 1 Schematic representation of scrap qualities and types.

energy needed to extract the ores from the natural environment.

Primary production involves metals produced from ore, whereas secondary production involves recycled metals or scrap. This term is not an expression of the quality of the final product. Recycled metals can come from three streams: home scrap, new scrap, and old scrap (postconsumer) (Fig. 1). Home scrap is produced in the metal-producing plant and is the result of inefficiencies or material losses in the plant. New scrap is scrap that comes from the manufacturing industry that purchased the metal. Both for home and for new scrap, the material composition is well-known, and therefore these types of scrap demand a higher price in the global scrap market. Old scrap is scrap recovered from a waste stream after consumption of the metal-containing products. The exact composition of old scrap is not known, which may limit recycling due to contamination with other metals that may be difficult or impossible to remove in the recycling process.

Metals and scraps are typically traded on a global scale. Hence, relative prices of primary metals and scrap are determined by global developments. For example, the economic crisis in Asia in the late 1990s triggered a decline in the world prices of steel scrap because Asia is a large net importer of steel scrap.

2. PRODUCTION OF METALS

Although society uses many metals, the following metals represent the bulk of the metals used: iron (steel), aluminum, copper, lead, and zinc. This article focuses on these five metals. Figure 2 presents the

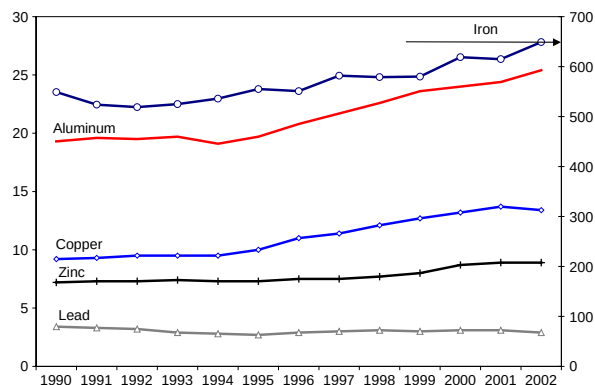


FIGURE 2 Global primary production of iron, aluminum, copper, zinc, and lead from 1990 to 2002. Iron production is depicted on the right-hand axis, whereas production of the other metals is depicted on the left-hand axis.

global primary production of the five metals. Iron (steel) clearly dominates the metal consumption on a global scale, representing approximately 90% of all metal consumption. All ores contain the oxide of the metal. The metal oxides are reduced to metallic form to be used by society. Because the chemical reduction process is endothermic, and because many reduction processes take place at high temperatures, it requires large amounts of energy. Besides a high energy intensity, the processes used to produce primary metals are also very polluting. Pollution may be due to the concentration of the ores, the smelting process (or processes producing inputs for the smelting, such as coke making to make iron), or energy generation.

Iron is the main metal component of steel. Global steel production exceeds 900 million tonnes, whereas global iron production was 649 million tonnes in 2002. Accounting for material losses in the processes, the difference between the amount of steel and iron produced is met by recycled scrap. Steel is produced in more than 60 countries. The main iron and steel producers are China, Japan, the United States, Russia, South Korea, and Germany. Iron ore contains on average approximately 58% iron. Iron is produced in two forms: pig iron (more than 93% of global iron production) and direct reduced iron (DRI). Pig iron is produced in a blast furnace using coke in combination with injected coal or oil to reduce sintered or pelletized iron ore to pig iron. Limestone is added as a fluxing agent. Coke is produced in coke ovens. Reduction of the iron ore is the largest energy-consuming process in the production of primary steel. DRI is produced by reduction of the ores below the melting point in small-scale plants (<1 million tonnes/year) and has different properties than pig iron. DRI production is growing

and approximately 5% of the iron in the world is produced by direct reduction, of which more than 90% uses natural gas as a fuel. Steel is made by reducing the carbon content of the iron. Primary steel is produced by two processes: the open-hearth furnace (OHF; 4% of world steel production in 2002) and the basic oxygen furnace (BOF; 60% of 2002 world steel production). The OHF is still used in different configurations, mainly in Eastern Europe, India, and China. The BOF process is rapidly replacing OHF worldwide because of its greater productivity and lower capital costs. In addition, this process needs no net input of energy and can even be a net energy exporter. Hot pig iron is mixed with up to 30% recycled scrap. The process operates through the injection of oxygen. Secondary steel is produced in an electric arc furnace (34% of 2002 world steel production) using scrap. Scrap is melted and refined using a strong electric current. DRI can be used to enhance product quality. Several process variations exist and fuels can be injected to reduce electricity use. The steel is then cast and shaped into different products. Traditionally, primary steel was the steel of choice to make thin and high-quality products such as hot and cold rolled sheets. Through improved quality control of the steel manufacturing process and technological development in casting and rolling, today secondary steel is also used to produce hot and cold rolled steel.

Aluminum is one of the most abundant metals in the earth's crust. Aluminum is found as bauxite, which contains approximately 25% alumina. Hence, for every ton of aluminum approximately 8 tons of bauxite is processed. Global aluminum production was more than 25 million tonnes in 2002. Primary aluminum is produced in 43 countries. The main producers are Russia, China, the United States, Canada, Australia, and Brazil. Bauxite is first concentrated in the Bayer process to produce pure aluminum oxide, or alumina. The Bayer process uses fuels and electricity. The alumina is then processed using the Hall-Heroult process. Used for more than 100 years, the Hall-Heroult process is an electrolysis process. The alumina is dissolved in molten cryolite and electrolyzed using carbon anodes. This is a very energy-intensive process, using large amounts of electricity (generally 13 or 14 MWh/tonne of aluminum). The production of carbon electrodes from tar pitch also consumes a considerable amount of energy. Because of the high electricity intensity, aluminum production is increasingly more concentrated in countries and areas with low electricity costs. Secondary aluminum is made by

melting and refining aluminum scrap. Because the scrap quality is often difficult to determine and to control, primary aluminum is typically used for high-quality applications such as thin sheet and foil, whereas secondary aluminum is mainly used for castings. Aluminum recycling rates vary for different countries depending on the primary aluminum products used. In the United States, approximately 3 million tonnes of aluminum scrap has been recovered, of which 60% came from new scrap and 40% from old scrap.

Copper has been the material of choice for electrical applications. However, competition from aluminum has reduced the role of copper in this market. Still, copper remains the third most important metal. Global primary copper production in 2002 was 13.4 million tonnes. Primary copper production is concentrated in less than 40 countries. The leading primary producers are Chile, Japan, China, and the United States. Copper ores that are currently mined have a very low concentration, typically less than 0.5%. This means that for every ton of copper more than 200 tons of ore is mined, transported, and processed. At the mine site, the ore is crushed and ground to concentrate the copper. The copper is extracted from the ore through leaching and other processes. The concentrate contains approximately 25% copper. The extracted copper oxide is then reduced in smelters at high temperatures to produce blister or anode copper. Different smelting processes are used throughout the world. The blister copper is further refined electrolytically. The smelting and refining process is highly energy-intensive. Secondary copper is produced by mixing some recovered scrap in the smelter, and it is then further refined. In 2001, only 12% of all copper was produced from secondary copper.

Zinc and lead are often found in the same ores. However, primary production is often separated. Zinc is the fourth most consumed metal in the world. Global primary production is estimated at approximately 9 million tonnes in 2002. Primary zinc is produced in 35 countries. The main producing countries are China, Canada, Australia, Japan, and the United States. Zinc is used for many applications and a relatively large portion of zinc consumption is for dissipative uses such as galvanizing steel and for chemicals. Zinc is mined from a variety of ores in coproduction with other metals. The zinc is leached from the ore and concentrated. The concentrate is often shipped to the zinc smelter. The zinc is smelted and refined in an electrolysis process using more than 4 MWh/tonne of zinc. Secondary zinc is often melted

and refined at separate plants and can also be recycled by primary zinc plants in the electrolysis process. Often, the secondary smelters produce zinc-based chemicals. In 2001, approximately 30% of the world zinc production was from secondary materials, including brass, die casting scrap, flue dust, galvanizing residues, and zinc sheet.

Lead is used for many applications, but the main market is lead-acid batteries (used in the automotive industry). Batteries represent approximately 90% of the market for lead. Lead production has remained stable for the past decade at approximately 3 million tonnes/year. Although demand for lead-acid batteries has increased, many other uses for lead have declined (e.g., as an additive in gasoline and paint) due to environmental and health concerns related to the dissipative use of lead. Lead is found in ores with cometals such as zinc. The lead ore is most often concentrated at or near the mine. Primary lead production is concentrated in 30 countries, with the largest producers being China, Australia, Canada, the United States, the United Kingdom, and Germany. The lead concentrate is smelted at high temperatures in a blast furnace. The product of the blast furnace is refined into the final lead metal. Lead is most often processed in secondary lead smelters. Most of the recycled lead is from old scrap recovered by specialized lead-recovery companies.

3. RECYCLING OF METALS

Recycling and recovery rates vary for the different metals. The recovery rate is defined as the amount of metal that is recovered after use, whereas the recycling rate is determined by the amount of metal actually recycled. These rates can be expressed as a function of apparent domestic consumption, domestic production, or the amount of discarded metal. Different researchers and countries use different system boundaries and definitions for both of these terms. Hence, it is difficult to easily compare the recycling rates between different countries and studies. Therefore, the following discussion focuses on available data for the United States collected and analyzed by the U.S. Geological Survey (USGS). The USGS expresses recycling rates as a function of apparent domestic consumption of the metal. Apparent consumption is the sum of domestic production and imports deducting exports of metal. The USGS includes both new and old scrap in the definition of recycled metals.

Globally, more than 370 million tonnes of iron and steel scrap is recycled by the iron and steel industry. In the United States, 71 million tonnes of steel scrap is recycled, compared to an apparent consumption of 119 million tonnes in 2001. The recycling rate is estimated to be 60%. Figure 3 presents the recycling rates for the major metals in the United States. Because steel production dominates the production of metals, the recycled amount of steel also dwarfs the volumes of other metals recycled. However, as Fig. 3 demonstrates, the relative recycling rate is also high compared to that of other metals. Only lead has a higher recycling rate. Unfortunately, no data are available on the contribution of old and new scrap for iron and steel. Because most steel is used in durable products (e.g., appliances, automobiles, and construction) as well as packaging, it is relatively easy to recover and recycle. Recycling rates of iron and steel vary by the metal market. It is estimated that 65% of all the steel in an automobile produced in the United States is recycled steel. However, the recycling rate of packaging cans is probably lower because not all communities have systems in place to collect and separate the steel cans. Note that tin can also be recovered from packaging cans in specialized detinning plants. After recovery of the tin, the steel is recycled in a steel plant.

Aluminum is recovered in almost all market segments in which aluminum is used. However, most secondary smelters focus on production of aluminum die casting for the automotive and appliance industry. There are no data available for the global amount of aluminum recycled. In 2001, almost 3 million tons of scrap was recovered in the United

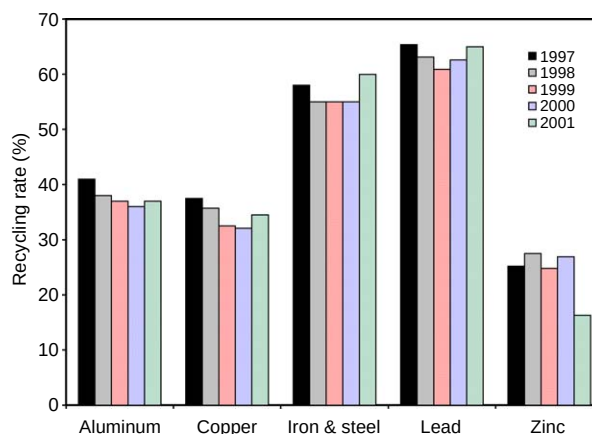


FIGURE 3 Recycling rates for selected metals in the United States for the period 1997–2001. Recycling rate is expressed as the amount of recycled metal as a share of total apparent consumption in the United States. From the U.S. Geological Survey.

States, representing a recycling rate of 37% (Fig. 3). Approximately 60% of the recycled scrap is from new (manufacturing) scrap and 40% from old scrap. In the United States, the major source of aluminum scrap is the aluminum beverage can, representing almost half of all old scrap. In 2001, 55.6 billion aluminum cans were recycled, which is equivalent to a recycling rate of 55% when compared to the manufacture of cans. Figure 4 depicts the relative shares of new and old scrap for the metals discussed in this article.

Global copper recycling was estimated to be 37% of the world copper consumption in 1997. In the United States, the total amount of recycled copper was estimated to be slightly more than 1.1 million tonnes in 2001. The recycling rate is relatively low (34.5%; Fig. 3) due to the dissipative use of copper as well as decreasing world market prices for copper. The recycling of copper concentrates on new scrap, whereas the recovery from old scrap has decreased in the past 5 years. Most copper is recovered from alloys and is mostly used by brass mills. New copper scrap represents more than 70% of all recycled copper.

Lead recycling demonstrates different dynamics than most other metals because most of the lead is recovered from postconsumer wastes (i.e., old scrap). This is due to the reduced dissipative use of lead and the concentration of lead consumption for automotive lead-acid batteries. In most industrialized countries, efficient and effective collection networks exist to recover the old batteries that are recycled by dedicated recycling companies. In the United States, approximately 96% of all recycled lead is from batteries and only 4% is recovered from new scrap (Fig. 4). In 2001, 1.1 million tonnes of lead was

recycled, whereas the total apparent consumption was 1.7 million tonnes. With a recycling rate of 65% (Fig. 3), lead is currently the most recycled metal in the United States.

Zinc use is dominated by dissipative uses such as galvanizing steel products. Hence, the recycling of postconsumer wastes of zinc is very limited. Zinc can be recovered from the waste of steel recycling. Zinc is recovered from the dust of electric arc furnaces (secondary steel production) and may be recovered from the sludge from primary steel-making plants. New technological developments aim at stripping zinc from steel. Hence, zinc recycling concentrates on the use of new scrap. This limits the net recycling rate of zinc. The declining recycling of old scrap in the United States has led to a reduced recycling rate of 16% (compared to as high as 28% in 1998), of which 86% is from new scrap (Fig. 4). Worldwide, approximately 30% of all zinc is produced from secondary materials.

This discussion of the five metals has highlighted the differences in recycling rates and the impact of the dominant uses on the recycling rates. However, it also shows that relatively large shares of (postconsumer) metal wastes are not recovered. This demonstrates the potential for future increases in recovery and recycling rates, which will be especially important because the production of home and new scrap will likely decrease due to improved material efficiencies of metal production and processing.

4. ENERGY USE AND RECYCLING

Metal production comprises approximately 7% of global energy use, making the sector one of the largest energy-consuming industries. Energy consumption is dominated by the production of iron and steel. An estimated 4% of world energy use is consumed in the production of iron and steel. Table I summarizes the estimated energy intensities for selected metals. The energy consumption data include the energy used in mining of the ores. It is important to note that there are relatively large differences in the energy intensity and efficiency of industrial processes in different countries and for plants within a country. Differences in energy use for metal production in various countries are due to differences in the raw material used (i.e., ore and scrap), the mix of products produced, material losses in the production process, and the energy efficiency of the production facilities. These four factors also contribute to continuous changes in energy intensity.

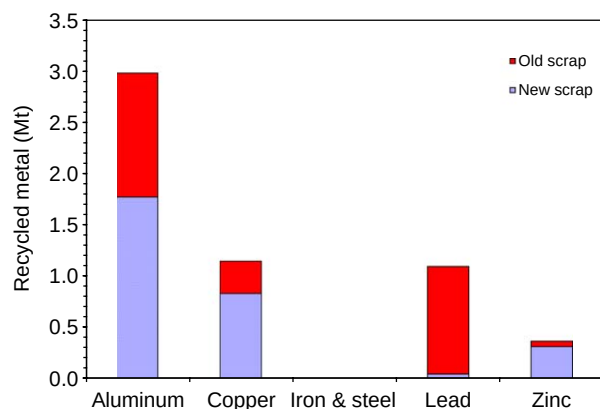


FIGURE 4 Recycled metal flows in the United States for 2001, subdivided into old (postconsumer) and new scrap. From the U.S. Geological Survey.

TABLE I

Energy Intensities of Primary and Secondary Metal Production^a

Metal	Primary production (million metric tons)	Energy intensity, primary production (GJ/tonne)	Energy intensity, secondary production (GJ/tonne)	Energy savings of recycling (%)
Iron and steel	649	26	12	54
Aluminum	25.4	213	11	95
Copper	13.4	91	13–37	60–85

^a Values are approximate and will vary by product and production location. Energy intensity is expressed as primary energy. Electricity is converted to fuels used to generate the power.

It has been estimated that the energy intensity in the U.S. iron and steel industry during the period 1980–1991 changed mainly due to improved energy and material efficiency and a change toward more recycling. However, in China, during the same period only energy efficiency improvement contributed to the change in energy intensity because it has less domestic scrap available. Good statistical information on energy use does not exist for all metal industries or countries.

Table I provides estimates for the energy intensity of primary and secondary metal production. The high energy use and energy intensity make this sector an important candidate for energy conservation policy. In metals production, energy can be saved through (i) improving energy efficiency of the production process, (ii) improving material quality and material efficiency, (iii) reuse of metal-containing products, and (iv) recycling.

Table I also provides estimates for the relative energy savings of recycling metals. The data demonstrate that the largest energy savings can be achieved in the recycling of aluminum. Energy savings in aluminum recycling is estimated to be 95%. For steel, the energy savings exceeds 50% of the energy inputs of primary steelmaking. These data demonstrate the importance of recycling strategies in an integrated energy and environmental policy framework.

Increased recycling will reduce the energy intensity of metal production. The net reduction in overall energy use in a country or the world will also depend on the growth in the demand for metals and other materials. Globally, there is strong growth in the consumption of all metals, which will likely offset the energy savings of recycling. However, without recycling the total energy needs for the increased mining and primary production would be much higher than currently expected.

Several countries have started to develop integrated energy policies that account for the potential energy savings of increased recycling and material efficiency efforts. Other countries have started to account for the energy savings as part of the rationale for recycling programs. Several studies have shown the benefits of sustained recycling programs and efforts to reduce future energy use and CO₂ emissions, highlighting the importance of integration of recycling in energy policies.

The energy consumption data in Table I exclude the energy needed to collect, recover, and transport the metals. Although considerable amounts of energy can be used to collect and recover postconsumer and old scrap, these are generally small compared to the energy savings. The efficiency and effectiveness of postconsumer programs will vary depending on the type of collection system operated, sustained lifetime, geography of the area served, as well as cultural characteristics. Based on the energy used in collecting and energy savings of recycling, an “energy optimal” recycling strategy can be designed. Such a strategy may well result in different optimal recycling rates based on the regional characteristics (Fig. 5). This optimum will change with changes in technology performance. A study of aluminum recycling in Germany showed that the current recycling rate of 59% could be improved to 79% to optimize energy savings from the recycling scheme using currently available technology.

Technology development will affect the energy consumption of primary and secondary production. New energy-efficient technologies are being developed that can substantially reduce the energy consumption in steelmaking through integration of different production steps (e.g., casting and shaping of steel products) and new designs of processes (e.g., electric arc furnaces with scrap preheating). For aluminum, new developments aim to reduce the

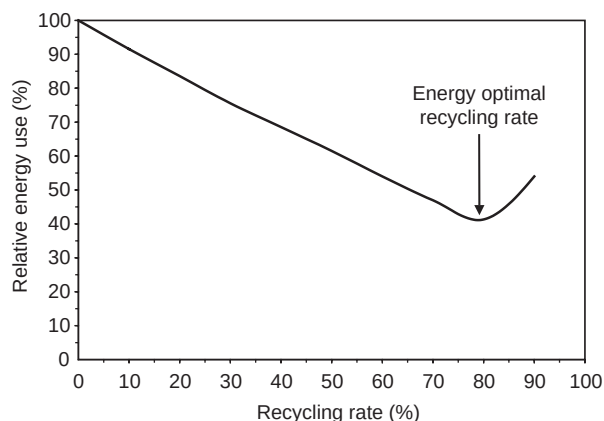


FIGURE 5 Schematic representation of energy use as a function of the recycling rate. The vertical axis depicts the amount of energy used for the production of a given volume of metal relative to the amount of energy needed to produce the primary metal (0% recycling is 100% of the primary energy use). Energy use includes the energy needed for collection, recovery, transport, and recycling of scrap.

electricity consumption of primary aluminum smelting through new electrode development and, in the long term, the development of totally new smelting processes to break through the thermodynamic barrier of the Hall-Heroult process. Secondary aluminum production is also improved through more efficient furnaces and improved yield of secondary melting furnaces. Similar developments are found for the other metals. New developments in metal recovery and separation technology will increase the recovery efficiency, improve quality control, and reduce energy intensity.

5. FUTURE OF RECYCLING

The future of metal recycling will be shaped by different driving forces. The main driving forces are depletion of high-quality ores, technology innovation, the availability of secondary resources, the costs of energy, and a drive toward more sustainable industry practices as part of sustainable development.

Depletion of high-quality ores has driven metal producers to extract ores with increasingly lower metal concentrations. Copper is an excellent example of the depletion of high-quality ore, such that ores currently exploited have concentrations less than 0.5%, whereas it was 2% at the beginning of the 20th century. This had led to exploitation of high-quality ore reserves at remote locations. Decreasing ore quality may not apply to all metals. For example, iron ore and bauxite reserves are abundant and are

found in many locations. However, there is competition to use the highest quality ore first. Also, geopolitical factors may play a role in the sourcing of raw materials and hence the costs of primary production. This process will increase the costs of primary metals and make recycling more attractive.

On the other hand, the increasing costs of ore extraction are also driven by technology innovation. Innovation and the introduction of new technology in the metals extracting and primary metals industry have reduced the cost of metal production over time. Therefore, although current copper ore contains on average only one-fourth of the copper that the ore contained a century ago, copper is now cheaper. Continued innovation in the metals industry will drive down the costs of primary production. However, the same applies to secondary metal production technology. Innovative processes are decreasing the costs of metal recovery and recycling while simultaneously improving the quality of the secondary metals. The rapid growth of the secondary steel industry in the 1990s in the United States is an example of how innovative technology and low resource (scrap) costs gave secondary steel producers a competitive edge over primary (integrated) steel producers. Within a decade, the secondary steel industry expanded to provide almost half of the steel production in the United States.

As previously discussed, the sources of scrap vary for the different metals. Most secondary metal producers prefer new scrap because the exact composition of the metal is known, enabling the production of high-quality secondary metals. However, due to innovations improving the energy efficiency of metal production and processing, the amount of new scrap will further decline in the future. Hence, future recycling efforts will focus more on old (postconsumer) scrap. Costs, quality, and availability will become more important factors with regard to secondary production. Innovative technologies may alleviate some of the quality and costs problems, whereas policies aimed at the reduction of dissipative use of metals and improved scrap recovery efficiency are needed to contribute to sustainable secondary metal production.

The high energy intensity of primary metals production makes this type of production much more vulnerable to energy price changes and supply disruptions. Already, primary production of the most energy-intensive metals such as aluminum takes place in countries with low-cost energy sources. The changing energy markets in the United States have contributed to the closing of primary aluminum smelters in the

northwest. The increased importance of climate change in energy policy worldwide will affect the costs and opportunity to use fossil energy. The primary metals industry will be affected more intensively by these changers than the secondary producers.

Finally, a drive toward sustainable development of our industrial society will affect the metals industry. Climate change is just one of the challenges for sustainable development because primary metals production is also a source of many other environmental pollutants. Historically, primary metal producers were major sources of acidic air pollution (from copper, nickel, and other metal smelters) and metal dust emissions. Primary metal production also results in large volumes of solid and fluid wastes (e.g., red mud from bauxite processing and sludge containing arsenic from metal leaching processes). A drive toward sustainable business practices will drive innovation but will also support further recycling efforts as well as material substitution and product changes. This will lead to an increased importance of metal recycling.

It is difficult to assess the contribution of the different drivers to the future of the metals industry. However, it is clear that metal recycling will play a more important role in the future, leading to further reductions in energy intensity in metal production.

SEE ALSO THE FOLLOWING ARTICLES

Aluminum Production and Energy • Material Efficiency and Energy Use • Material Use in Automobiles

• Recycling of Paper • Reuse and Energy • Steel Production and Energy

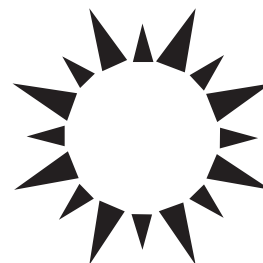
Further Reading

- Ayres, R. U. (1997). Metals recycling: Economic and environmental implications. *Resour. Conservation Recycling* 21, 145–173.
- Fenton, M. D. (1998). “Iron and Steel Recycling in the United States in 1998.” U.S. Geological Survey, U.S. Department of the Interior, Reston, VA.
- Melo, M. T. (1999). Statistical analysis of metal scrap generation: The case of aluminium in Germany. *Resour. Conservation Recycling* 26, 91–113.
- Michaelis, P., and Jackson, T. (2000). Material and energy flows through the UK iron and steel sector, Part 1: 1954–1994. *Resour. Conservation Recycling* 29, 133–158.
- Quinkertz, R., Rombach, G., and Liebig, D. (2001). A scenario to optimise the energy demand of aluminium production depending on the recycling quota. *Resour. Conservation Recycling* 33, 217–234.
- Ruth, M. (1998). Dematerialization in five U.S. metals sectors: Implications for energy use and CO₂ emissions. *Resour. Policy* 24, 1–18.
- Tilton, J. E. (1999). The future of recycling. *Resour. Policy* 25, 197–204.
- U.S. Geological Survey (2001). “Recycling—Metals.” U.S. Geological Survey, U.S. Department of the Interior, Reston, VA.
- Worrell, E., van Heijningen, R. J. J., de Castro, J. F. M., Hazewinkel, J. H. O., de Beer, J. G., Faaij, A. P. C., and Vringer, K. (1994). New gross energy-requirement figures for materials production. *Energy Int. J.* 19(6), 627–640.
- Worrell, E., Price, L. K., Martin, N., Farla, J. C. M., and Schaeffer, R. (1997). Energy intensity in the iron and steel industry: A comparison of physical and economic indicators. *Energy Policy* 25(7/8), 727–744.

Recycling of Paper

LINDA GAINES

Argonne National Laboratory
Argonne, Illinois, United States



-
1. Introduction
 2. Production and Recycling Statistics
 3. Primary Paper Production
 4. Use of Recovered Paper
 5. Conclusion

Glossary

chemical pulp Pulp obtained by digestion of wood with solutions of various chemicals. Chemical pulps are used to make shipping containers, paper bags, printing and writing papers, and other products requiring strength.

corrugated container A box manufactured from containerboard, layers of linerboard, and one layer of medium. The layers are combined and corrugations pressed into the medium.

de-inking A process in which most of the ink, filler, and other extraneous material are removed from recovered paper, resulting in a pulp that can be used in the manufacture of new paper.

direct energy Fuel or electric energy consumed during a process.

indirect energy Energy required to produce a material or chemical input to a process.

kraft paper Paper made from wood pulp by a modified sulfate pulping process. Its natural unbleached color is brown, but it can be bleached to white. It is used as a wrapping paper or converted into bags, envelopes, or coated paper.

kraft (sulfate) pulp A strong papermaking fiber produced by the kraft chemical process in which the active cooking agent is a mixture of sodium hydroxide and sodium sulfide. The term kraft is derived from a German word meaning strong.

mechanical pulp Wood pulp manufactured by mechanical processes. The resulting paper is opaque and has good printing properties, but it is weak and discolors easily when exposed to light due to residual lignin. Uses include newsprint, specialty papers, tissue, toweling, paperboard, and wallboard.

newsprint A lightweight paper, made mainly from mechanical wood pulp, engineered to be bright and opaque for the good print contrast needed by newspapers.

paper The name for all kinds of matted or felted sheets of fiber (usually vegetable, but sometimes mineral, animal, or synthetic) formed on a fine screen from a water suspension.

primary fiber Pulp made from trees or other vegetable matter rather than recovered materials. Also called virgin fiber.

pulp Fibrous material prepared from wood, cotton, grasses, etc. by chemical or mechanical processes for use in making paper or cellulose products.

pulp substitutes A combination of paper and paperboard trimmings from paper mills and converting plants.

recycled or secondary fiber Cellulose fiber reclaimed from waste material and reused, sometimes with a minor portion of virgin material, to produce new paper.

After paper products have served their intended uses, there are several alternative paths for material disposition. These include (generally, but not necessarily, in order of decreasing energy saved) reuse, recycling to the same product, recycling to a lower valued product, combustion for energy recovery, incineration without energy recovery, and landfill. The best option depends on the decision criterion selected.

1. INTRODUCTION

After paper products have served their intended uses, there are several alternative paths for material disposition. These include (generally, but not necessarily, in order of decreasing energy saved) reuse, recycling to the same product, recycling to a lower valued product, combustion for energy recovery, incineration without energy recovery, and landfill. Solid-waste management is an important issue in the United States, and the priorities are a subject of

debate. According to the U.S. Environmental Protection Agency (EPA), "Recycling (including composting) is the preferred waste management option to further reduce potential risks to human health and the environment, divert waste from landfills and combustors, conserve energy, and slow the depletion of nonrenewable natural resources." Combustion for energy recovery is generally not counted toward material recovery goals, probably because it is viewed as a form of disposal and is therefore assumed to waste resources and have negative environmental effects. However, cocombustion in coal-fired plants or combustion in appropriately pollution-controlled waste-to-energy (WTE) plants is safe, avoids landfill costs, and can displace fossil fuels. In some cases, more fossil fuels can be displaced by combustion than by recycling. This article evaluates alternative product life cycles and discusses trade-offs among possible goals. For instance, office paper recycling conserves trees but may require greater fossil fuel input than virgin paper production. Technical and institutional barriers could arise as recycling rates increase. The main areas of concern are energy and materials use and landfill volume, but key economic and environmental factors are also identified.

2. PRODUCTION AND RECYCLING STATISTICS

In the period from 1990 to 2000, production of paper and paperboard in the United States increased from 80.5 to 96.4 million tons. This represents a 6% per capita increase, from 644 to 683 pounds per person per year, despite predictions that increased use of electronic communications would move us toward a "paperless society." Since most paper products have very short use times, postconsumer paper waste generated during the same period tracked production closely, increasing from 72.7 to 86.7 million tons. Paper and paperboard have accounted for approximately 37% by mass [and approximately an equal percentage of volume because they have close to the average municipal solid waste (MSW) density] of MSW during the period, according to EPA statistics.

This discarded material represents a valuable resource and a costly liability to discard, and efforts to utilize it have increased. The percentage of postconsumer material recovered from 1990 to 2000 doubled from 20.2 to 39.4%. Note that the

paper industry reports higher recovery numbers because it includes in-house scrap, such as pulp substitutes, and materials such as unsold magazines that are recycled immediately without actually being used, as well as postconsumer material. More paper is recovered from MSW streams than all other materials combined. Corrugated boxes (recovered at a rate of 71%) account for 54% of all paper recovered, and newspapers (with a 58% recovery rate) account for another 22% of recovered paper. Office papers (of which 54% are recovered) account for another 10%; no other category contributes as much as 5%. Therefore, energy and materials implications of paper recycling are of special interest.

A significant quantity of wastepaper is exported. Some is mixed waste, sent to low-labor-cost countries where it can be economically sorted by hand, but higher grade materials are also included. Export of high-grade wastes could adversely affect recycled fiber quality and has implications for energy use. Similarly, imports have energy use implications. In general, imports of finished products displace domestic resources, but they worsen the trade balance. The costs and benefits of imports and exports of finished products and scrap must both be considered.

3. PRIMARY PAPER PRODUCTION

More than 48% of the raw material fiber used for papermaking in the United States in 2002 was recovered or reclaimed, with recovered paper contributing 37.7% and wood residues 10.6%. The total has been more or less constant, with the percentage of recovered paper increasing and that of wood residues declining recently due to their increased use in oriented strand board and composite panels. Primary fibers are a renewable, domestic resource, much of which is currently supplied by tree plantations. Short-rotation forestry can be used to further increase the productivity of these plantations. Therefore, use of primary fibers need not mean cutting old-growth natural forests, with the implied loss of habitat, wilderness, and recreation land. However, monoculture may have environmental implications of its own, as does recycling; these should be taken into account during decision making about paper production and recycling policy.

For purposes of this article, virgin paper can be classified into two main types that have different chemical compositions and physical properties. In reality, the industry produces a large variety of products, but these two classes cover most of them.

The two types are distinguished by the processes used to make the pulp: chemical or mechanical. Kraft pulp (the most common type of chemical pulp) and thermomechanical pulp (TMP), plus recycled fiber (chemical or mechanical), are the largest inputs to U.S. papermaking (85% of total pulp input), and their use is growing. This article focuses on these types. The energy and materials inputs to production of the different types of pulp differ, although both start with wood as the basic raw material. Papers made from different types of pulp are often kept separate in recycling because of the differences in material properties, but the secondary processing sequences are similar.

Most discussions of energy use in papermaking are confusing because of a failure to define exactly what is meant by energy. It is important to keep track of the form in which energy is used. The major inputs to a paper mill are wood, purchased energy in the form of electricity and fossil fuels, and water. Some components of the wood are used as a raw material and some as by-product fuel. What the industry generally reports is the process energy use at the mill, which is the sum of the purchased energy and the process by-product energy (part of the energy originally in the wood) that is used. It is more useful to define total energy input to include the entire energy content of the wood as well as the purchased fossil fuels and electricity because the wood input could otherwise have been used as a fuel. Some of this energy is not used up; it is still available from combustion of the final product. (The difference between the total energy input and the residual energy content of the product is equal to the process energy as tracked by the industry.) Total energy input, purchased and process energy, and the net nonwood energy consumption for typical paper production and recycling are shown in Table I for discarding, burning, and recycling 1 ton of printing and writing paper ("office" paper) and newsprint (representing typical products from chemical and mechanical pulps). The table shows that total energy input is reduced by recycling, but that nonwood energy use is actually increased by recycling of office paper. Recycling office paper may actually require more fossil fuel purchases than producing it from virgin materials because the recycling process does not generate by-product fuels and must therefore use purchased energy.

Caution is in order regarding energy consumption in paper mills. The numbers cited are representative of typical U.S. industry practice for specific grades and can be used to provide an understanding of the factors

involved in recycling decisions. However, actual paper mill energy use varies widely, depending on many factors, including the age of the equipment, the product and its degree of bleaching, the geographical location, and the season. For instance, recycling of paperboard requires less energy than recycling of office paper and may therefore be preferable to combustion. It is hoped that the American Forest and Paper Association will publish actual industry data that will allow more accurate analysis. In addition, different mills use different mixes of purchased and self-generated electricity, affecting both the fuel mix and the primary fuel usage for generating electricity. Also, the fuel mix for electricity generation varies from utility to utility, affecting the actual primary energy usage from electricity purchases (or the fuel displaced if electricity is sold back to the grid). Table I was prepared using the U.S. average generation mix, but decisions for specific areas should employ actual local utility data.

Energy and resources used for papermaking in other areas of the world could be quite different from those used in the United States. Therefore, the conclusions of this article may not necessarily be applicable to other regions. For example, in Europe, where wood resources are more limited, clay fillers are used more extensively in the production of newsprint to produce a lighter weight, less wood- and energy-intensive product. Nonwood plant fiber pulp is used extensively in China, whereas very little is used in the United States and Europe. Nonwood plants are fast growing and produce a different fiber than slower growing trees. Many nonwood fibers are short like hardwood fibers, but others are much longer and must be shortened mechanically or chemically to optimize their papermaking value. Process energy requirements differ with fiber characteristics.

3.1 Chemical Pulping

Chemical pulping separates wood chips into their components of cellulose, hemicellulose, and lignin and retains the cellulose fibers while rejecting most of the rest; therefore, the yields are less than 50%. However, much of the material removed is used for process fuel, so energy purchases are low. The two main types of chemical pulp are kraft (sulfate) and sulfite, indicating the types of chemicals used. Most printing and writing papers, packaging, and paperboard are made from kraft pulp because of its superior strength and economics. Sulfite pulp has a relatively small and declining market share and will not be discussed further. The total direct process

TABLE 1

Energy Required to Supply 1 Ton of Paper to Consumers^a

Path	Wood ^b (dry tons)	Heat of combustion of wood (10 ⁶ Btu)	Purchased fuel (10 ⁶ Btu)	Purchased electricity (kWh)	Process energy ^c (10 ⁶ Btu)	Energy recovered (10 ⁶ Btu)	Energy displaced (10 ⁶ Btu)	Net total energy input (10 ⁶ Btu)	Nonwood energy input (10 ⁶ Btu)
Wood to kraft ^d to landfill	2.2	37.4	9.5	495	36.5	0	0	52.1	14.7
Wood to kraft to WTE	2.2	37.4	9.5	495	36.5	15.6	10.9	52.1	3.8
Recycled fiber to kraft recycling	0.33	5.6	7.4	656	14.3	0	0	19.9	14.3
Wood to news to landfill	1.05	17.9	0	3079	32.3	0	0	50.2	32.3
Wood to news to WTE	1.05	17.9	0	3079	32.3	16	11.9	50.2	20.3
ONP to news to recycling	0.21	3.6	5.0	1436	20.1	0	0	23.7	20.1

Source. Gaines and Stodolsky (1993).^a Abbreviations used: WTE, waste to energy; ONP, old newsprint.^b Wood for kraft paper includes bark, which is used as by-product fuel. Newsprint input excludes bark, generally used elsewhere in integrated mills.^c Electricity converted at 10,500 Btu/kWh; values are lower if some of the electricity is cogenerated.^d Kraft refers to printing and writing grades; other paper or board types would have slightly different energy requirements.

energy requirement reported by the industry for pulping and making of printing and writing papers averages approximately 26×10^6 Btu/ton, but the energy required varies greatly with the product and with the degree of energy recovery included, ranging from a high of approximately $30\text{--}35 \times 10^6$ Btu/ton for kraft board to a low of $10\text{--}12 \times 10^6$ Btu/ton for recycled tissue. Process energy use is generally higher for products requiring extensive bleaching. Using paper that has undergone less bleaching could reduce fuel purchases more than recycling. There is a public misconception that virgin paper is white and recycled paper is brown. However, it is the degree of bleaching that determines brightness, so virgin paper can be brown and recycled paper can be white. Reducing bleaching also reduces environmental impacts. The primary purchased energy requirement for kraft paper, over half of which is bleached, is typically approximately 15×10^6 Btu/ton.

3.2 Mechanical Pulping

Products made from mechanically produced fibers include newsprint, high-bulk news (pocket books), coated publication paper, and chipboard. Newsprint is generally made from TMP, although some other mechanical pulps are still used, and some kraft pulp may be added to increase strength. The energy required to make 1 ton of newsprint (both direct and indirect energy) is all purchased (no process by-product energy) and ranges between 27.6 and 39.4×10^6 Btu/ton. This range primarily results from

variation in pulping energy and electrical power-generation efficiency. It is assumed that all steam needed by the process comes from the waste heat generated in the TMP process. Therefore, these estimates include electrical power only. The mid-range estimate is 33.5×10^6 Btu/ton, which is calculated on the basis of typical pulping energy and 50% in-plant power generation.

4. USE OF RECOVERED PAPER

4.1 Paper Recycling

Wastepaper is pulped mechanically and possibly de-inked before bleaching and papermaking. These upgrading processes, used to recycle printing and writing papers, remove ink and any coatings or fillers that were used in the paper. They are reasonably energy-intensive, using approximately 2×10^6 Btu/ton (excluding energy embodied in chemicals). However, they are essential for upgrading of recycled material for utilization in high-grade products, such as printing and writing paper. High-grade recycled products therefore may have higher energy requirements than shown in this analysis and yield lower energy savings. Segregated wastepaper streams such as paperboard, newsprint, or corrugated generally do not require de-inking.

For de-inked fiber of modest quality, the yield is typically approximately 67%. This means that each ton of wastepaper yields 0.67 tons of fiber

and 0.33 tons (dry basis) of sludge, which is sent to the landfill. Sludge is difficult to dewater and at best only approximately 50% solidifies, resulting in a mass to the landfill of 0.66 tons. Thus, for every ton of de-inked fiber, 1 ton of sludge goes to the landfill or is available for combustion. Although the Office of Technology Assessment classified this sludge as noncombustible, it can be gasified or dried with stack gas and then burned. The sludge has a heating value of approximately 14×10^6 Btu/ton (dry basis) from the fiber and ink content. There was previously concern about toxic metals from printing inks in the sludge (or in the ash from sludge or MSW combustion); however, the inks that contained these materials have been phased out in the United States.

The production of secondary paper results in fiber loss. Therefore, virgin fiber is needed to maintain a constant supply of paper to the consumer. Very little newspaper is reported to be rejected at the materials recovery facility, but 15–22% of the fiber input to the mill is lost in the recycling process because of fiber breakage. There is no way to avoid this loss. Thus, the recovery rate for collected material is 78–85%. The energy needed to produce the primary pulp that makes up for the loss must be included in the energy needed to recycle. Direct energy purchases by a modern secondary newsprint mill are typically approximately 16×10^6 Btu/ton of recycled paper product; the total primary energy needed, including transportation and embodied energy in chemicals, is approximately 20.8×10^6 Btu/ton. The energy needed to supply newsprint to the consumer, assuming 22% fiber loss, is thus 23.6×10^6 Btu/ton ($0.78 \times 20.8 \times 10^6$ Btu/ton for the recycled material plus $0.22 \times 33.5 \times 10^6$ Btu/ton for the makeup material).

For recycled paper products, all of the energy is generally purchased, although some mills do generate some of their own energy by combustion of sludge. This can reduce energy purchases by as much as a factor of two. Technically, if paper is recycled within an integrated mill, some of the energy used may have been generated as a by-product of other processes within the mill. However, it cannot be credited to the recycling process and should still be considered as imported or purchased.

Recycling processes may also differ by region. Most European and Japanese mills started using flotation technology for de-inking before U.S. mills did so. However, at least one major U.S. mill still uses washing to de-ink the recovered fiber. The flotation process became dominant in Europe mainly because of effluent regulations (washing utilizes more water) and the higher recovery yields of

finer and fillers (the high levels of clay fillers in European paper necessitate high levels of recovery to maximize yield).

4.2 Fiber Energy Recovery: Use of Wastepaper as Fuel

Energy can also be saved by burning waste newsprint or kraft paper (instead of recycling it) and recovering energy as electrical power or steam. This option should be called fiber energy recovery and is not a type of disposal. It should not be confused with incineration, in which wastes are burned but no energy is recovered. Approximately 10% of old newsprint (ONP) in the United States is removed from the waste stream and utilized for fiber energy recovery. The energy generated (normally electrical power) displaces energy produced from fossil and nuclear fuels at central-station power plants. When 1 ton of paper is burned in a WTE plant, 11.3×10^6 Btu of energy is displaced. If paper (news or kraft) is pelletized (either alone or as part of mixed waste) for cocombustion in coal-fired power plants, the generation efficiency of electricity produced from the waste is that of the power plant, rather than the lower efficiency of the WTE plant, and more energy is displaced. In cocombustion, each ton of paper displaces 16.8×10^6 Btu of coal. In addition, combustion can be accomplished in an environmentally sound manner; emissions of sulfur oxides (SO_x) are actually reduced by cocombustion of waste pellets, so more costly pollution controls can be avoided.

Wastepaper can be burned in either existing woodwaste- or coal-fired boilers or in new, efficient boilers designed to burn recovered fiber/wastepaper. Depending on the type of boiler, it may be necessary to pelletize the paper fuel to ensure clean and complete combustion. The energy produced can displace fossil fuel use in pulp and paper mills, cement kilns, and other industrial or commercial facilities. If electricity is cogenerated, any excess power generated can be sold back to the grid, thereby displacing additional fossil fuels. Although total energy use may not be reduced by this strategy, renewable biomass energy is substituted for fossil fuels, and the result is a reduction in emissions of CO_2 as well as SO_x , NO_x , and volatile organic compounds. In addition, the volume of solid waste is reduced. Pelletization of mixed waste paper (MWP) requires the addition of moisture; one manufacturer supplies this via paper recycling sludge, which is otherwise a disposal problem.

This technology has been used in at least one large paper mill; the boiler technology is commercially available. The boilers should be no more costly than those firing fossil fuels, although retrofits and pelletizing equipment could be needed and add to total costs. An economic factor that could hinder use of this technology is the extremely volatile price of wastepaper. This is, in turn, related to the varied mandates for recycled content in paper. Recognition of fiber energy recovery as an acceptable form of recycling could help reduce this barrier. One potential drawback to combustion for energy recovery is that it could result in competition with paper recycling mills for feedstock. This could drive up the price of recovered paper and jeopardize the economic viability of both industries.

A key institutional and economic factor is whether the paper to be burned is classified as a solid waste. If it is so defined, more stringent emission requirements may apply to the boiler in some states, and additional, costly pollution control measures may be required, negatively impacting the process economics. Requirement of additional controls is inappropriate because combustion of wastepaper actually reduces boiler emissions; therefore, either used paper should be classified as fiber fuel (with less stringent regulations) and not as a solid waste or facilities switching to it should be exempted from equipment installations specified in waste-combustion new source performance standards (NSPS) or from the "Best Available Control Technology" review on fuel switching if they can demonstrate that they have reduced their emissions to levels at or below those of the NSPS.

Compared to production of recycled office paper using fossil fuel, production of virgin office paper using wastepaper as the fuel would reduce CO₂ emissions by approximately 2 tons per ton of paper, in addition to reducing fossil fuel use. The result is that paper is produced with zero or negative net CO₂ impact. This is because biomass is a renewable resource. It is assumed that trees cut down to produce paper will be replanted and again sequester the CO₂, so there is no net CO₂ impact on the environment. The sulfur content of paper is much lower than that of the coal or oil it displaces; SO_x emissions may be reduced by more than expected by simply displacing coal or oil with paper because of reactions with compounds in the paper. NO_x emissions are also reduced, as is the quantity of solid waste generated.

Chesapeake Paper Products Company (no longer in business) estimated energy use and emissions for a boiler that would have burned recovered fiber at its

pulp and paper mill in West Point, Virginia, as well as for an efficient new mill. In both cases, combustion of recovered fiber would have reduced emissions and fossil fuel use. Georgia Pacific Corporation cofired wastepaper pellets with low-sulfur coal in a water-tube stoker boiler at its mill in Jarratt, Virginia. It compared steady-state operation of the boiler with 100% coal feed to that with 30% coal and 70% wastepaper pellets (by weight); measured emissions of CO₂, CO, SO₂, NO_x, and hydrocarbons; and analyzed the ash. Results were very favorable.

In addition, feasibility studies were completed of combustion of MWP at three facilities in Washington State: a paper mill, a hospital, and a cement kiln. Shredded MWP could be fired in a cement kiln without the need to pelletize it, with no additional controls required because of the nature of the existing process. Hospitals and paper mills would require additional pollution controls if the paper were classed as a waste. Economics were found to be favorable if disposal costs were avoided. MWP was obtained from recycling brokers for proximate and ultimate analysis, and it was found to have acceptable fuel characteristics, but insufficient funding was available for test burns. However, data from a test burn of 50% coal/50% pellets at the University of Wisconsin showed reductions in emissions of SO₂, NO_x, and metals compared to 100% coal. Additional feasibility studies with favorable results were carried out in Canada, Oregon, and Wisconsin. A small college in Iowa has used pellets to supply much of its steam.

4.3 Effects of Recycling

The maximum purchased energy that could be saved by recycling newspapers in the United States, assuming a constant supply to the consumer, is 150×10^{12} Btu/year or the equivalent of approximately 8 million tons of coal annually ($[33.5 \times 10^6 \text{ Btu/ton} - 23.6 \times 10^6 \text{ Btu/ton}] \times 15.0 \times 10^6 \text{ tons}$). The savings at current levels of recycling are lower. Similar energy savings could be achieved by burning ONP and recovering energy that would displace fossil fuels in electricity generation. However, if newspaper is recycled, the trees that would have been used for replacement newspaper are also saved. Therefore, because trees are a valuable resource, recycling of newspaper makes sense and should be encouraged. The purchased energy difference between recycling and WTE options for newspaper is

small, but the total energy input, including wood, is reduced by recycling of newsprint.

However, the situation for office paper is very different. In a modern kraft mill, paper can be made from wood with little purchased energy because of the use of by-product fuels, whereas recycling requires significant energy purchases because no fuel by-products are produced when wastepaper is pulped. Therefore, a decision to recycle office paper is a decision to use fossil fuels in place of wood. Recycling paper does keep it out of landfills, but so does combustion in a WTE or power plant, and energy recovery displaces additional fossil fuel. Combustion of paper in MSW (excluding newsprint) could displace more than 650×10^{12} Btu/year of fossil and nuclear fuels $[(60 \times 10^6 \text{ tons}) \times [15.6 \times 10^6 \text{ Btu/ton}] \times [0.7, \text{ the relative efficiency of WTE plants to utilities}]]$. This is equivalent to approximately 30 million tons of coal annually or approximately 3% of U.S. annual coal consumption.

The best option for paper depends on the objective. In the case of office paper, there is a choice. To save fossil energy and reduce landfill volume, burn wastepaper. To save trees, recycle paper. Therefore, it may be inappropriate for waste management planners and policymakers to exclude combustion for energy recovery from the acceptable waste disposition alternatives, as proposed U.S. legislation did very explicitly. Combustion for energy recovery (as opposed to incineration) as an appropriate option to recycling has been accepted in Europe, however, with The Netherlands adopting a 40% goal and Sweden 60%.

Significant energy conservation could be achieved by reuse of some paper products, such as corrugated boxes. Newspapers could also be reused. There are bins located in subway stations for collection of newspapers. The message on the bins is "Read, Ride, Recycle." They could also include "Reuse" in their list. Newspapers left by early commuters could be read by those arriving later. One possible mechanism would be to hire some of the people already soliciting money to reassemble and sell complete used papers at a reduced price. This would benefit the worker and the environment to a much greater extent than would recycling the paper after one use.

Another factor to be considered in decisions regarding paper production and disposition is emission of CO_2 to the environment, which is of concern because of possible global warming. Growing trees respire (take in) CO_2 at a rate that varies with their age. The new trees in the plantations that replace the ones cut for fiber take in more CO_2 than do the old

forests. A study for Electric Power Research Institute (EPRI) estimated the effects on CO_2 emissions of various materials options for paper production and disposition. This study concluded that producing paper from plantation trees and burning the wastepaper for energy recovery is preferable, from the standpoint of greenhouse gas emissions, to all other options, including recycling. Thus, the same option for disposition of office paper is chosen if greenhouse gas reduction is the objective as when fossil fuel use and old-growth forest cutting are to be minimized.

Because a large amount of wastepaper is exported, the energy implications of used paper exports are considered briefly. For most categories of paper, U.S. consumption exceeds production, so net exports of waste imply an increase in primary production for domestic use. The balance of trade is better served by exporting higher value-added finished products. Export of old newsprint without importing newsprint increases domestic energy use and cutting of trees and does not serve the goal of energy conservation. Although export does avoid landfill, so would combustion or recycling, which would recover the embodied energy domestically. For office paper, an increase in primary production over secondary production implies greater wood use but smaller fossil fuel purchases. Therefore, if domestic energy conservation is the primary goal, it may be appropriate to encourage exports of waste office paper but not of old newspaper.

5. CONCLUSION

The most beneficial product life-cycle depends on the goals to be achieved. Trade-offs may have to be made, as in the case of kraft paper, in which trees and fossil fuels are traded off in the recycling decision. Therefore, the objectives of decision makers must be well defined before appropriate recommendations can be made. The optimal mix of reuse, recycling, and combustion for the United States may not necessarily correspond to the optimal mix for Europe and other areas of the world. Reuse is often overlooked but may be the best option when possible.

Technical and institutional barriers to increased recycling rates must be considered before mandating high recovery levels. High rates mean either separation of mixed waste, with its attendant technical difficulties, or development of improved source-separation procedures. The success of source-separation programs depends on the willingness of the

public to cooperate. Such incentives as bottle deposits have had some success; the limits of public cooperation need to be explored. In general, the cleanest, most easily recognizable waste products are recovered first. As recovery rates rise due to government mandates, economic pressures, or possibly public awareness, the quality of the material recovered will decline, and the energy and costs required to produce an acceptable product will rise. Eventually, a break-even point will be reached, at which the benefits of recycling are outweighed by the cost. This break-even point depends on the material and on the measures of cost being used.

Recycling of newspaper saves trees (a material and an energy resource) and fossil fuels, whereas combustion only displaces fossil fuels; recycling is preferable. Production of paper from chemical pulps, on the other hand, depends primarily on wood inputs for production but on fossil fuels for recycling, and the product can be burned to displace fossil fuels. More energy could be recovered by combustion of office paper than by recycling. In addition, kraft paper represents approximately 70% of the volume of material discarded, so it must be kept out of landfills. Combustion and recycling prevent paper from filling up landfills. A life cycle for office paper that included virgin production from dedicated plantations (to save old-growth forests and their value for recreation and wildlife habitat) and product combustion for energy recovery would minimize fossil fuel use as well as landfill volume. This is, in essence, substitution of biomass for fossil fuels. On the other hand, a life cycle including recycling minimizes the use of trees. The best answer depends on the grade of paper, the location, the fuel mix for the plants and the electricity supply, and the details of the process design. All of the alternatives must be evaluated objectively before decisions are made.

Acknowledgments

I thank the many helpful people in the paper industry who provided real-world feedback on the information in this article, enabling important updates and clarifications. Thanks especially to Benjamin A. Thorp at Georgia Pacific, who provided information, references, and personal contacts. Those who provided particularly useful input include Tom Brown at Carastar, Jim Ramp at SP Newsprint, and Dedric Bergstrom at Glatfelter.

SEE ALSO THE FOLLOWING ARTICLES

Material Efficiency and Energy Use • Recycling of Metals • Reuse and Energy

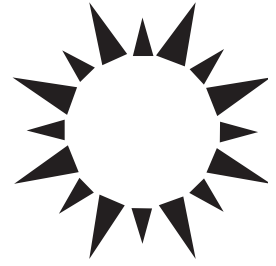
Further Reading

- American Forest and Paper Association web site: <http://www.afandpa.org>.
- Butner, R. E., Kangas, M., and Stapley, C. E. (1993). Paper recycling: Impact on electricity-use, electro-technology opportunities, Report RP-3228-06. H. A. Simons Strategic Services Division for Electric Power Research Institute, Decatur, GA.
- Environmental Defense Fund, Duke University, Johnson & Johnson, McDonald's, The Prudential Insurance Company of America, and Time Inc. (1995, December). White paper No. 3, lifecycle environmental comparison: Virgin paper and recycled paper-based systems. (Updated February 2002).
- Gaines, L. L., and Stodolsky, F. (1993, December). Mandated recycling rates: Impacts on energy consumption and MSW volume, Report ANL/ESD-25. Argonne National Laboratory, Argonne, IL.
- Office of Solid Waste and Emergency Response, U.S. Environmental Protection Agency (2002, June). Municipal solid waste in the United States: 2000 facts and figures, Report EP1530-R-02-001. Office of Solid Waste and Emergency Response, U.S. Environmental Protection Agency, Washington, DC.
- U.S. Department of Commerce, Bureau of the Census (2002). "Statistical Abstract of the United States: 2002," 122nd ed. U.S. Department of Commerce, Washington, DC.

Refrigeration and Air-Conditioning

BOGDAN D. HORBANIUC

Technical University of Iași
Iași, Romania



-
1. Introduction, Applications, and Historical Background
 2. Basic Principles of Refrigeration
 3. Reversed Thermodynamic Cycles
 4. The Air–Standard Refrigeration Cycle
 5. Vapor–Compression Refrigeration Cycles
 6. Absorption Refrigeration Systems
 7. Steam Jet Refrigeration
 8. Other Methods of Refrigeration
 9. Cryogenics
 10. Air-Conditioning

Glossary

air conditioner System composed of a refrigeration machine, heat exchangers, air ducts, filters, fans, and/or a humidifier that supplies air at specified parameters (temperature, humidity, and purity).

air-conditioning The simultaneous control of the temperature, humidity, and cleanliness of the air in a space as required either by the thermal comfort of persons that live or work in the respective space or by specified technological prescriptions.

Carnot cycle Thermodynamic cycle having theoretically the highest efficiency; used as a criterion in evaluating the effectiveness of a real cycle.

coefficient of performance Thermodynamic efficiency of a refrigeration cycle, expressed as the ratio between the heat removed from the low-temperature body and the external work consumed in order to remove that heat and to transfer it to the high-temperature body (surroundings).

cryogenics The deep-refrigeration domain involving temperatures less than 120 K.

refrigerant Substance in the fluid state that transfers to the surroundings the heat removed from the cooled space.

refrigeration The creation and maintenance within a medium of a temperature lower than that of the surroundings.

refrigeration system Complex aggregate of interrelated machines, heat transfer equipment, and control devices that together accomplish the refrigeration task.

The second law of thermodynamics, as stated by Clausius, asserts that heat can be transferred from a body of low temperature to a body of higher temperature only if external work is expended. The high-temperature body is called cooled medium (space) and the fluid that carries the removed heat and transfers it to the surroundings is called refrigerant. To be able to accomplish these tasks, the refrigerant must perform a sequence of thermodynamic processes the effect of which is the refrigeration of the cooled medium. The temperature domains of refrigeration correspond to air-conditioning (approximately 5°C), industrial refrigeration (0 to –150°C), and cryogenics (–150 to –273°C).

1. INTRODUCTION, APPLICATIONS, AND HISTORICAL BACKGROUND

The applications of refrigeration cover almost all domains of human activity: the food industry (processing, storage, and transportation of aliments), air-conditioning, ice manufacture, chemical industry (removal of the heat of reaction or of the heat of mixing, salts separation from solutions, and gases liquefaction), civil engineering and mining (freezing of moisture in unstable water-bearing soils), the manufacturing industry (heat treatments and metal cutting), the pharmaceutical industry, biomedical applications (blood and tissue storage and hypothermia in surgery), and fundamental research (atomic and nuclear physics).

In 1755, William Cullen obtained ice by evaporating water under low pressure. In 1805, Oliver Evans described theoretically the operation of the refrigerant's vapor compressor. J. Leslie developed the first absorption refrigeration system using the water-H₂SO₄ solution in 1810. In 1823, Michael Faraday published the first classical work on gases liquefaction and liquefied SO₂, H₂S, CO₂, and NH₃. Sadi Carnot described the refrigeration cycle in 1824. The first patent on a mechanically driven refrigeration system belongs to Jacob Perkins (1834, London), who proposed ether as a refrigerant. In 1844, John Gorrie, an American physician, constructed the first compression refrigeration machine using air as refrigerant. The first viable commercial system was put into operation in 1857 (James Harrison and D.E. Siebe) using ethyl as a refrigerant. The first absorption refrigeration machine was patented by Ferdinand Carré in 1859 and was designed to use the water-ammonia solution. At the beginning, these machines were used to produce ice. More than 600 absorption refrigeration machines had been sold in the United States by 1876. A. C. Kirk patented the reversed Stirling air refrigeration cycle in 1861. Carl von Linde (Germany) developed the first air liquefaction machine in 1895, and in 1908 Kammerlingh Onnes (Holland) liquefied helium. The first primitive version of a household refrigerator dates from 1898. In 1910, Maurice Leblanc (France) constructed the first steam jet refrigeration machine. Migley, Henne, and McNary (United States) used chlorofluorocarbons (CFCs) for the first time as refrigerants in 1930, and in 1933 the adiabatic demagnetization of paramagnetic salts (theoretically developed in 1926 by Debye and Giauque) was applied by W. F. Giauque and D. P. MacDougall to obtain temperatures of approximately 0.7 K. The lithium bromide-water solution was utilized after 1940 in industrial applications of the absorption refrigeration cycle (the Carrier and Trane systems). Small refrigeration machines that work on the basis of the reversed thermoelectric effect have been in use since 1958 to cool electronic components. In the 1980s, the adiabatic demagnetization of the nuclear spin system of copper allowed a temperature of 10⁻⁸ K to be reached.

The applications of refrigeration in the area of food processing and preservation have given impetus to the economic development of the meat-, fruits-, and vegetables-producing countries (usually located in the warm climate zone), allowing them to export these goods long distances. The evolution of the perishable foodstuff refrigeration techniques was

consequently strongly related to the development of these countries. In 1873, Charles Tellier presented to the Academy of Sciences a note regarding fresh meat refrigeration. In 1875, the ship *Paraguay*, equipped according to Carré's specifications, transported refrigerated meat at -30°C from Buenos Aires to Le Havre, France. In 1876, Tellier uses for the first time an absorption refrigeration system to preserve meat, fruits, and vegetables on the ship *Le Frigorifique* during its journey from Rouen, France, to Buenos Aires. In 1882, Argentina used for the first time refrigeration systems in slaughterhouses.

Refrigerated fruits were marketed for the first time in the United States in 1910. In approximately 1930, refrigerated vegetables were marketed for the first time. Refrigeration trucks have been in use in the United States, Italy, and Great Britain since 1933. In 1938, the Hoveman quick refrigeration method (direct contact under vacuum of the foods contained in latex sacks) was developed.

2. BASIC PRINCIPLES OF REFRIGERATION

The refrigerant passes through the refrigeration system by continuously removing heat from the cooled medium and rejecting this heat to the surroundings (atmospheric air). To accomplish its task, the refrigerant must be in a fluid state. Due to the fact that the latent heat of vaporization-condensation is much higher than the sensible heat, the change of state of the refrigerant is preferred in order to maximize the transferred heat (the refrigeration effect). Consequently, in most refrigeration systems the cooling effect is achieved by the evaporation of a liquid, a process that extracts the amount of heat necessary to vaporize the fluid. By using a fluid that boils at a low temperature, heat is removed from the respective body (space) by cooling it. The greater the removed heat (latent heat of vaporization), the more intense the cooling effect. The refrigeration processes that occur in such a manner that the refrigerant undergoes a single time the heat removal from the cooled medium are called open-cycle processes. Some examples of open-cycle processes are (1) Water evaporation in atmospheric air: water cools the air by removing from it an amount of heat equal to the latent heat of evaporation, and (2) Endothermic mixtures: by using substances that require heat to mix, the formation of the mixture removes the heat from the space that

contains the mixture. Such endothermic mixtures contain different percentages of water and salts such as sodium chloride, ammonium chloride, calcium chloride, and ammonium carbonate.

Refrigeration processes that occur when the refrigerant, after undergoing a sequence of physical transformations, regains its initial state and repeats the same sequence of transformations a number of times are called closed-cycle processes. The basic refrigeration techniques (vapor-compression refrigeration, absorption refrigeration, and steam jet refrigeration) are examples of closed-cycle processes.

2.1 Methods of Refrigeration

Thermodynamic methods

- Open-cycle processes (evaporation of a liquid and refrigerant mixtures)
- Closed-cycle processes (mechanical vapor-compression, absorption refrigeration, steam jet refrigeration, and expansion of a compressed gas in centrifugal field of forces)

Electrical and magnetic techniques

- Reversed thermoelectric effect
- Adiabatic demagnetization of a paramagnetic salt

Methods used in cryogenics

- Gas throttling (the Joule-Thomson effect)
- Expansion of a gas with work output
- Reversed Stirling cycle
- Linde method
- Claude method
- Diffusion pump method.

The refrigeration capacity has been defined as the heat rate necessary to freeze a short ton (907 kg) of ice in 24 h. One ton of refrigeration = 12,000 Btu/h = 3.51 kW.

3. REVERSED THERMODYNAMIC CYCLES

Thermodynamic cycles allow the reciprocal conversion in a continuous manner of thermal energy into other kinds of energy. A work-supplying cycle is referred to as the power cycle (Fig. 1). Such cycles are traced clockwise and they remove heat from the hot reservoir, supply work, and reject heat to the cold reservoir. A reversed cycle is traced counterclockwise and consumes external work in order to cool the cold reservoir by transferring heat to the high-temperature reservoir (Fig. 1).

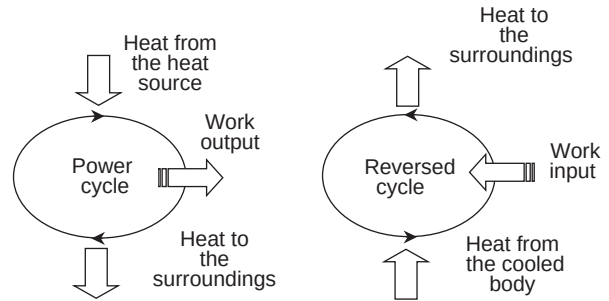


FIGURE 1 Power and reversed cycles.

Reversed cycles are used either to extract heat from bodies colder than the surroundings (cooling effect), referred to as refrigeration cycles, or to heat spaces by extracting heat from low-temperature sources (heating effect), called heat pump cycles.

The efficiency of a cycle is evaluated by the coefficient of performance (COP). In general, the effectiveness of a cycle is the ratio of the desired effect to the energy expended to obtain this effect. In the case of power cycles, the desired effect is the net work transfer, and the expended energy is the heat extracted from the high-temperature reservoir. For the reversed cycles, the desired effect depends on the purpose of the cycle, whereas the energy expense is the same, namely the external work input:

$$\text{COP} = \frac{\text{desired (useful) effect}}{\text{work input}}$$

If the purpose is the refrigeration of a space, then the desired outcome is the refrigeration effect (i.e., the heat removed from the low-temperature body):

$$(\text{COP})_{\text{refrig}} = \frac{\text{refrigerating effect}}{\text{work input}}$$

The actual COP of the refrigeration systems is on the order of 3.

On the other hand, if the purpose of the cycle is to heat a space, the heating effect (the heat rejected to the high-temperature body) represents the required result:

$$(\text{COP})_{\text{heat pump}} = \frac{\text{heating effect}}{\text{work input}}; (\text{COP})_{\text{heat pump}} = 4-5.$$

The reversed Carnot cycle is the reversible (ideal) cycle that has the thermodynamically highest possible COP and is composed of two isothermal and two isentropic (adiabatic) processes. Figure 2 shows the Carnot refrigeration (a-b-c-d-a) and the Carnot heat pump (1-2-3-4-1) cycles traced in the temperature-entropy plot. The processes that compose

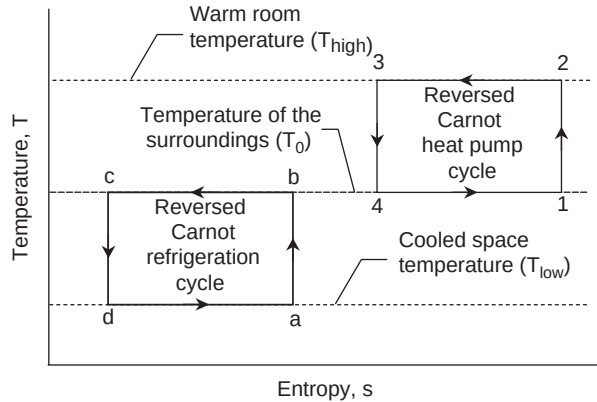


FIGURE 2 Reversed Carnot cycles.

the Carnot refrigeration cycle are the adiabatic (isentropic) compression a–b from T_{low} to T_0 , the isothermal heat rejection b–c to the surroundings at T_0 , the adiabatic (isentropic) expansion c–d from T_0 to T_{low} , and the isothermal heat removal d–a from the cooled space at T_{low} . The COP of such a cycle is

$$(\text{COP})_{\text{refrig}}^{\text{Carnot}} = \frac{1}{(T_0/T_{\text{low}}) - 1}$$

The Carnot heat pump cycle consists of the adiabatic (isentropic) compression 1–2 from T_0 to T_{high} , the isothermal heat rejection 2–3 to the warm room at T_{high} , the adiabatic (isentropic) expansion 3–4 from T_{high} to T_0 , and the isothermal heat removal 4–1 from the surroundings at T_0 . The COP of a Carnot heat pump cycle can be expressed as

$$(\text{COP})_{\text{heat pump}}^{\text{Carnot}} = \frac{1}{1 - (T_0/T_{\text{high}})}$$

Perfect isothermal processes cannot be obtained with gases as working fluids since such processes require either infinitely low rates of expansion or compression to maintain the temperature rigorously constant at a finite heat transfer rate or infinitely high rates of heat transfer at finite rates of expansion or compression. The only way to obtain the isothermal processes of a reversed Carnot cycle is to use a fluid as refrigerant that liquefies from the gas state by ejecting its latent heat of vaporization at the higher temperature and that vaporizes at the lower temperature of the cycle by removing from the cooled space an amount of heat equal to the latent heat of vaporization. This cycle is represented in Fig. 3.

There are two major drawbacks to the reversed Carnot cycle with phase-change refrigerant:

1. The adiabatic compression between T_{low} and T_{high} occurs in the liquid–vapor region, which is why

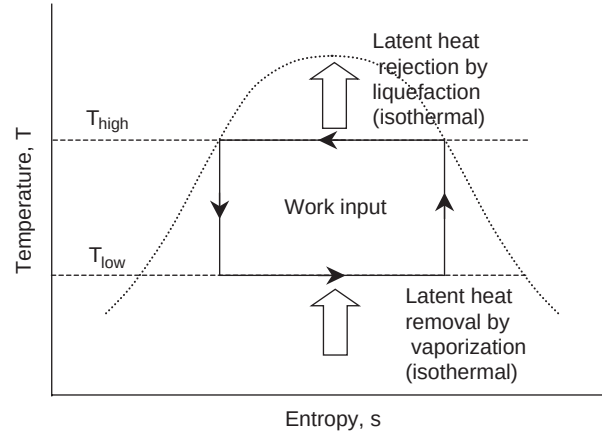


FIGURE 3 Reversed Carnot refrigeration cycle with liquefaction and vaporization of the refrigerant.

it is called wet compression. Due to the presence of the liquid phase, the liquid droplets may damage the compressor blades (erosion and pitting).

Solution: The compression should occur in the superheated vapor region.

2. The expansion involves a low-quality liquid–vapor mixture and therefore the turbine’s efficiency is extremely low. Although this process provides work for the compressor, the price to be paid in terms of efficiency and difficult operation conditions does not justify the presence of the expander–turbine. The advantage of the net work production is surpassed by the losses.

Solution: The expansion process in the turbine is sacrificed in favor of a more thermodynamically effective process: throttling. The expander is replaced by a throttling valve, which destroys the potential energy of pressure and reduces the refrigerant’s pressure from p_{high} to p_{low} but makes the system more reliable and simple.

The result is the so-called vapor–compression refrigeration cycle.

4. THE AIR-STANDARD REFRIGERATION CYCLE

The adiabatic expansion of a gas is accompanied by a significant decrease in temperature. Therefore, this phenomenon can be used in refrigeration cycles. When refrigerant leakage may occur in the cooled space occupied by living beings, or when air-conditioning is required (such as in the case of aircraft), air is the recommended working fluid. Other advantages result from the fact that air is available in unlimited quantities and involves no

manufacturing costs. Due to the low COP of the gas refrigeration systems, their use is recommended only when very well justified.

The air-standard refrigeration cycle is the reversed Brayton cycle. The direct Brayton cycle is the basic cycle of operation of gas turbine power systems.

4.1 Reversed Brayton Cycle

The cycle is composed of two isentropic processes (compression and expansion), an isobaric heat rejection to the surroundings, and an isobaric heat removal from the cooled space.

The reversed Brayton cycle requires relative high pressure ratios and large volumetric flow rates to achieve modest refrigeration capacities. For usual applications, vapor-compression or absorption refrigeration systems are more suitable from the standpoint of refrigeration capacity per system size. However, by modifying the scheme gas refrigeration systems can be used to liquefy air. Use of a regenerative heat exchanger, which cools the air below the temperature of the surroundings, leads to an important temperature decrease after the expansion, which means that the cooling effect is much more intense.

Figure 4 shows the scheme of the system and the temperature-entropy plot of the cycle. The processes that comprise the cycle are the adiabatic compression (1–2) of the air from p_{low} to p_{high} (in the compressor), the isobaric heat rejection (2–3) to the surroundings, the adiabatic expansion (3–4) of the air from p_{high} to p_{low} (in the turbine), and the isobaric heat removal (4–1) from the cooled space. The turbine supplies a fraction of the necessary work input for the compressor.

The functions of the environmental control system of the aircraft are to adjust the cabin temperature, pressure, and humidity; cool avionics; defog the windshield; and supply conditioned air to various systems and subsystems. The heat that is removed from the previously mentioned parts of the aircraft is transferred to the ambient air and to the vehicle's fuel supply. Figure 5 shows the schematic of the air refrigeration system for aircraft cooling and the cycle's T - s (temperature-entropy) diagram.

Atmospheric air is absorbed by the compressor, and it heats up and rejects heat to the surroundings. The cooled air expands under the turbine and its temperature decreases under the prescribed cabin temperature. The cooled air leaving the expander is mixed with the warm cabin air, and the result is refreshed air with the imposed temperature. In addition, a fraction of the cooling air is directed to the avionics and other systems that require cooling. The dashed line in Fig. 5 represents a fictitious process necessary to close the cycle.

The COP of the cycle strongly depends on the pressure ratio $p_{\text{high}}/p_{\text{low}}$: The lower the ratio, the higher the COP. Consequently, the compression ratio is limited to the values in the range 3–5. On the other hand, this fact leads to a decrease of the refrigeration effect (the heat extracted from the cooled body). To increase this amount of heat, regeneration is required: The air leaving the cooled space cools the air that leaves the heat transfer surface of the hot end of the refrigeration system. The air leaving the cooled space (state 1) is cold enough to cool the stream of air that has rejected heat to the ambient (state 3). Consequently, a regenerative heat exchanger is placed before the expander in order to achieve this supplementary cooling.

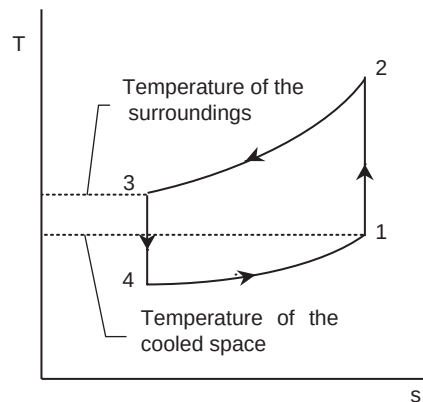
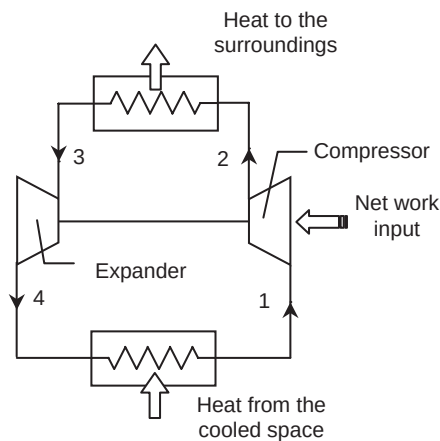


FIGURE 4 Scheme of the typical reversed Brayton refrigeration system and the temperature-entropy diagram of the cycle.

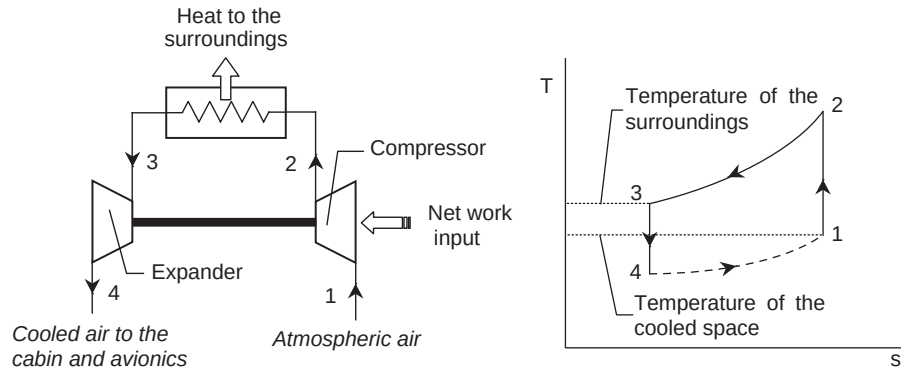


FIGURE 5 Open-cycle arrangement of the reversed Brayton refrigeration system for aircraft cooling.

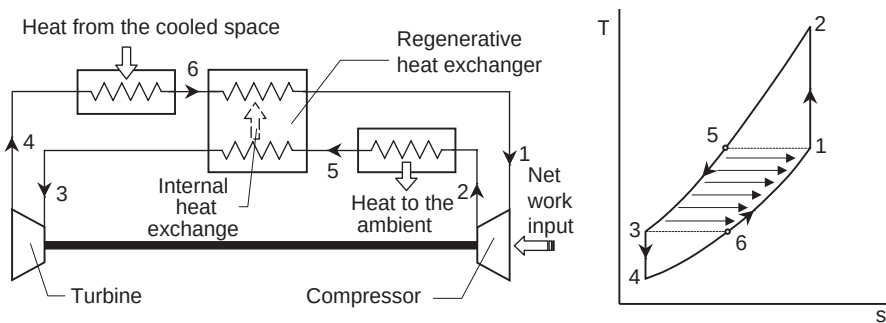


FIGURE 6 The air cycle with heat regeneration.

Figure 6 shows the reversed Brayton refrigeration system with heat regeneration. The processes are basically the same as for the reversed Brayton cycle. In addition, an internal heat exchange occurs in the regenerative heat exchanger: The air that rejected heat to the surroundings undergoes a supplementary cooling (process 5–3) by rejecting heat to the air that leaves the cooled space (process 6–1).

In order to reduce the work required by the system and thus to increase the COP, compression may be divided into two stages with intercooling. The pressure ratio for the two compression stages (compressors) must be the same so as to minimize the work input.

4.2 Bootstrap System

When using ram air as a heat sink to reject the generated heat within an aircraft, the corresponding takeoff gross weight (TOGW) penalty can be defined as the sum of the ram air circuit equipment weight and the additional fuel. Optimizing the aircraft performance means minimizing the ram air circuit TOGW penalty, and consequently the most lightweight and simple systems are required. This is the case for air-standard refrigeration systems. Due to

severe weight limitations, most of the aircraft cooling systems use the open-loop bootstrap configuration.

Open-loop bootstrap systems use bleed air from the aircraft's engine compressor. This stream is compressed by the bootstrap compressor, cooled by the ram air, and passes through a turbine that drives the bootstrap compressor. The temperature of this stream after passing through the turbine is lower than the temperature of the load and thus cooling is performed and heat is rejected overboard together with the air.

Figure 7 shows a schematic of an open-loop bootstrap system with heat regeneration. Air is compressed in the aircraft engine compressor from the ambient pressure to the combustor pressure. In the lower stages of the engine compressor, a fraction (called bleed air) is extracted at an intermediate pressure and is sent to the bootstrap compressor, where its pressure and temperature are increased. Next, the air stream passes through the regenerative cross-flow heat exchanger, where it undergoes a preliminary cooling by rejecting heat to the “used” air that leaves the cabin. In the ram air heat exchanger, the ram air coming from the surroundings cools the air. In the bootstrap turbine, air is expanded to the cabin pressure and to a temperature lower than

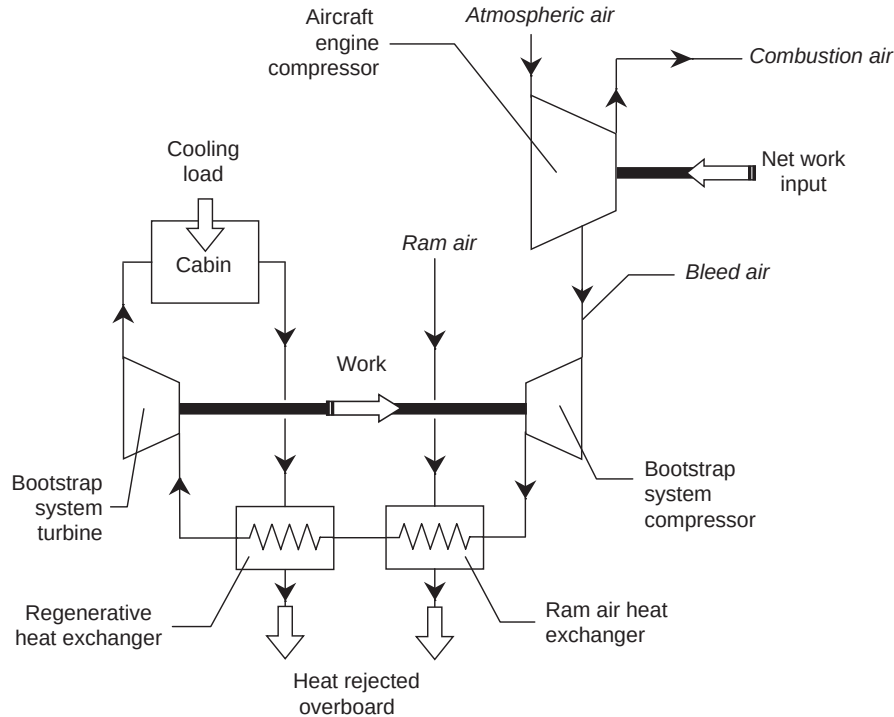


FIGURE 7 Open-loop bootstrap system for an aircraft.

the imposed cabin temperature. By flowing through the cabin and the aircraft systems, it removes the cooling load and afterward is sent to the regenerative heat exchanger to cool the incoming air. Finally, the used air is thrown overboard, as is the ram air. In the open-loop bootstrap system, the power supplied by the bootstrap turbine equals the shaft work rate absorbed by the bootstrap compressor.

5. VAPOR-COMPRESSION REFRIGERATION CYCLES

As discussed previously, the major drawbacks of the reversed Carnot cycle with vaporization and liquefaction make the Carnot refrigeration machine inappropriate for practical applications. By replacing the turbine (expander) with a throttling device and moving the compression into the superheated vapor domain, these shortcomings are eliminated and the result is the vapor-compression refrigeration (VCR) cycle.

5.1 Basic Vapor-Compression Refrigeration Cycle

The four basic components of a VCR system (Fig. 8) are discussed here. The compressor increases the

refrigerant's pressure from p_{low} to p_{high} (the isentropic process 1–2 in Fig. 8) and consequently its temperature becomes higher than that of the high-temperature region. In the condenser, the refrigerant rejects to the surroundings (or to a cooling medium) its latent heat of condensation and liquefies (the isobaric process 2–3). The expansion (or throttling) valve reduces the refrigerant's pressure (the isenthalpic process 3–4) from p_{high} to p_{low} , which is characterized by a boiling point below the temperature of the cooled space. In the evaporator, the refrigerant boils by removing heat from the refrigerated space (the isobaric process 4–1) and turns into superheated vapor that is sucked by the compressor. This process is repeated (Fig. 8).

The condenser and the evaporator are heat exchangers and the expansion valve is a capillary tube or a short-tube orifice. The throttling process occurs at constant enthalpy, which is why it is called isenthalpic. In modern refrigeration systems, an active throttling device that senses the temperature in the evaporator replaces the capillary tube and adjusts the pressure drop in order to control the cooling effect.

The expansion of the high-pressure liquid in the expansion valve results in a low-quality, two-phase fluid. The COP of the vapor-compression

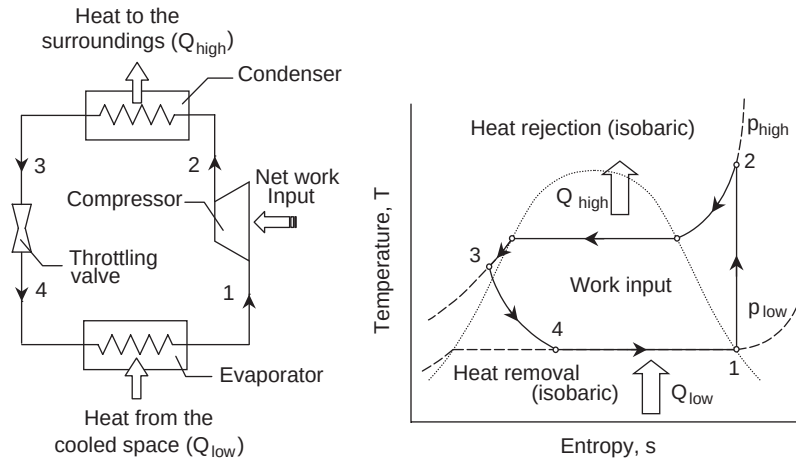


FIGURE 8 Scheme of the basic vapor compression refrigeration system and the cycle representation in the temperature-entropy diagram.

refrigeration system is given by

$$\text{COP} = \frac{h_3 - h_1}{h_2 - h_1}$$

where h is the specific enthalpy.

Irreversibilities inevitably occur in actual processes, and therefore the actual performance of the vapor-compression refrigeration system is less effective than that of the ideal cycle. Such an irreversible process is the nonisentropic compression in the compressor, which results in larger work input and an additional vapor superheating that increases h_2 and thus decreases the COP. Another effect that reduces the COP (h_3 decreases) is the slight liquid supercooling after the condensation due to the variation in temperature of the surroundings.

The lowest cooling temperature of the system is determined by the pressure ratio of the compressor and by the characteristics of the refrigerant. In a simple cycle, typical cooling temperatures vary from -50 to 2°C . Reciprocating compressors operate with pressure ratios up to 9:1. Outside the previously mentioned temperature range, either the compressor performance becomes unacceptably poor or the risk of excessive heating endangers lubrication, bearings, etc. By staged compression, the inferior temperature limit is lowered to -100°C . There are limitations regarding the refrigerant's characteristics: the high specific volume at low temperatures, which necessitates large size equipment; the too small specific volume of the condensed liquid; and the too high condensing pressure. In this case, the solution is to use cascaded systems with two or more refrigerants, each of which operates in a separate cycle. The advantage of this approach is the possibility to

choose for each cycle the best refrigerant from the standpoint of properties, which ensures the best possible operation. With cascaded systems, the lowest temperature that can be obtained is approximately -150°C .

Although centrifugal compressors may operate with pressure ratios up to 18:1, due to the low efficiency caused by the high temperature at the end of compression, the practically acceptable pressure ratio is the same as that for reciprocating compressors. To increase the efficiency, interstage cooling is used in the so-called economizers. In such equipment, relatively cool gas is mixed with the hot gas resulting from the compression process.

5.2 Refrigerants

More than 30 refrigerants are known, of which 10 are the most used. Refrigeration involves a large variety of applications and parameters. Consequently, one refrigerant cannot meet all the requirements. This is why there are different refrigerants.

5.2.1 Requirements

The general criteria to be met by a refrigerant candidate are

- High latent heat of vaporization in order to minimize the amount of refrigerant used
- Freezing point below the operating temperature at the cold end of the system
- Condenser pressure sufficiently low to avoid too high mechanical stresses in the components of the refrigeration system

- Evaporator pressure above the atmospheric pressure in order to eliminate the risk of leakage of air and moisture into the refrigeration system
- Critical parameters (temperature and pressure) well above the operating ones
- Low specific volume in the suction section of the compressor; large specific volume requires a larger size compressor and high cross section areas of the low-pressure piping
- Low manufacturing cost
- Good heat transfer properties (thermal conductivity, specific heat, and viscosity)
- Nonflammable, nontoxic, noncorrosive, and nonexplosive.

The toxicity of refrigerants, combined with their flammability, is rated by a class system. Class A designates refrigerants that are nontoxic at concen-

trations less than 400 parts per million (ppm) and class B refers to refrigerants that are toxic at less than 400 ppm in the air. Odorless toxic refrigerants should contain a warning agent (e.g., acrolein, which is an irritant to eyes). The flammability is evaluated by numbers ranging from 1 to 3 (1, nonflammable; 2, moderately flammable; 3, highly flammable). For instance, a refrigerant rated A2 indicates that it is nontoxic at less than 400 ppm and moderately flammable. Ammonia is rated B2.

5.2.2 Classification

The refrigerants are classified as showed in Table I. Almost half of the vapor-compression refrigeration systems use ammonia as refrigerant. Other working fluids usually referred to as halocarbons or halogenated hydrocarbons are the CFCs and the hydrochlorofluorocarbons (HCFCs), which are generally

TABLE I
Classification and Physical Properties of the Refrigerants

Class	Chemical name	Chemical formula	Number designation	Normal boiling point (°C)	Critical temperature (°C)	Critical pressure (bar)
Inorganic compounds	Water	H ₂ O	—	100	374	221.29
	Ammonia	NH ₃	R-717	-33.35	132.4	113
	Carbon dioxide	CO ₂	R-744	-78.5	31	73.3
	Sulfur dioxide	SO ₂	R-764	-10.01	157.5	78.8
Hydrocarbons	Methane	CH ₄	R-50	-161.5	-82.5	44.4
	Ethylene	C ₂ H ₄	R-1150	-103.5	9.5	51.2
	Ethane	C ₂ H ₆	R-170	-88.6	32.2	48.9
	Propylene	C ₃ H ₆	R-1270	-47.7	91.4	46
	Propane	C ₃ H ₈	R-290	-42.3	96.8	42.6
	Isobutane	C ₄ H ₁₀	R-600	-11.7	113.7	36.7
	Tetrachloromethane	CCl ₄	R-10	76.7	283.14	45.6
Chlorofluorocarbons	Trichlorofluoromethane	CCl ₃ F	R-11	23.7	214.1	34.1
	Dichlorodifluoromethane	CCl ₂ F ₂	R-12	-29.8	112	41.15
	Chlorotrifluoromethane	CClF ₃	R-13	-81.5	28.8	38.6
	Tetrafluoromethane	CF ₄	R-14	-128	-45.5	37.5
	1,1,2-Trichlorotrifluoroethane	CCl ₂ CFClF ₂	R-113	47.7	214.1	34.1
	1,2-Dichlorotetrafluoroethane	CClF ₂ CClF ₂	R-114	3.5	145.7	32.8
	Monochloropentafluoroethane	CClF ₂ CF ₃	R-115	-38	80	32.4
Hydrochlorofluorocarbons	Dichloromonofluoromethane	CHCl ₂ F	R-21	8.9	178.5	51.66
	Chlorodifluoromethane	CHClF ₂	R-22	-40.8	96	49.36
	Dichloromethane	CH ₂ Cl ₂	R-30	39.2	235.4	59.7
	Chloromethane	CH ₃ Cl	R-40	-24	143	66.8
	Chlorodifluoroethane	CH ₃ CClF ₂	R-142	-9.25	136.45	41.5
	Trifluoroethane	CH ₃ CF ₃	R-143	-47.6	73.1	37.8
	Difluoroethane	CH ₃ CHF ₂	R-152	-25	113.5	44.9
Azeotropes mixture	R-22, R-115	—	R-502	-45.6	90	42.6

nonflammable and nontoxic. CFCs and HCFCs have proven to be harmful to the ozone layer and contribute to the greenhouse effect responsible for global warming. For this reason, national and international regulations have restricted their use, recommending their substitution with other fluids that are not harmful.

Hydrocarbons such as propane and butane are cheap and suitable as refrigerants, except they are flammable. However, careful handling, a good system design, and the avoidance of leakage by appropriate safety measures reduce or eliminate the risks of explosion.

Mixtures of two or more refrigerants are also used. They are of two types: zeotropic mixtures, which exhibit a saturation temperature range (called temperature glide) at constant pressure, and azeotropic mixtures, which behave like a single component, evaporating or condensing at a fixed temperature at constant pressure (Table I).

5.2.3 Environmental Problems, Regulations, and Substitution Agents

In 1986, world production of CFCs was 950,000 tons. The highest CFC quantities are used in air-conditioning (house and automotive); in the United States and Japan, 40 and 25%, respectively, of R-12 production in 1986 were directed to this area of applications. Automobile production in 1990 was 40 million pieces (40% air-conditioned).

In 1974, Rowland and Molina discovered that CFCs destroy the ozone layer, which acts like a shield for life on Earth against the cosmic ultraviolet (UV) rays. In 1985, a 40% depletion of the ozone layer was observed. The mechanism of the phenomenon, which takes place in the stratosphere, consists of the reaction (triggered by UV radiation) of a molecule of CFC with ozone (O_3), the result of which is the liberation of an atom of chlorine, which combines with ozone to form oxygen and a molecule chlorine-oxygen. This molecule may react with a free atom of oxygen, resulting in molecular oxygen and a free chlorine atom, which again reacts with ozone and the process is repeated.

The ozone-depletion potential (ODP) of a refrigerant defines the potential damage caused by the refrigerant molecule to the ozone in the stratosphere and is evaluated based on the estimated life of the refrigerant in the atmosphere. It seems that ozone depletion is proportional to the squared concentration of chlorine in the atmosphere. This concentration is approximately 3 parts per billion in volume, and it is estimated that it will double by 2020. The life span in

the upper atmosphere is 120 years for R-12, 90 years for R-113, and 65 years for R-11. CFC molecules are nontoxic and chemically stable (inert) in the lower atmosphere and in the troposphere (up to 10 km in altitude). They become chemically active in the stratosphere, where they are decomposed by the cosmic UV radiation and are destroying the ozone layer.

Another harmful effect of the CFCs and HCFCs relates to the fact that they are opaque to the infrared radiation emitted by the earth as a result of its exposure to the sun's thermal radiation. As a result, infrared thermal radiation is confined in the system Earth-atmosphere and thermal energy is accumulated instead of being emitted, leading to global warming. The previously described mechanism constitutes the greenhouse effect (the glass covering a greenhouse acts in a similar manner by heating the space of the greenhouse).

The intensity of this phenomenon associated with a CFC or HCFC is evaluated by means of the halocarbon global warming potential (HGWP). For instance, a molecule of R-12 has the same effect as 5000 molecules of CO_2 .

CFCs exhibit the highest ODP and HGWP. HCFCs are less stable and therefore their ODP and HGWP are correspondingly lower. Hydrofluorocarbons (HFCs) have zero ODP since fluorine does not react with ozone, but some of them are characterized by equal or larger HGWP than that of the HCFCs. Hydrocarbons have zero ODP and HGWP. Figs. 9 and 10 illustrate the relative ODP and HGWP of some of the most popular CFCs.

Due to their harmful effects, regulations have been issued by governments in order to ban or to reduce the use and production of these substances. In the

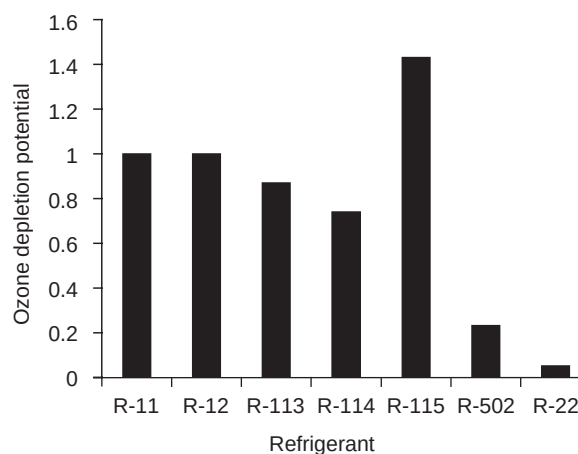


FIGURE 9 The relative ozone-depletion potential of some of the most used CFCs (normalized with respect to R-11).

late 1970s, the United States and Canada prohibited the use of CFCs as aerosols. In 1987, the Montreal Protocol was signed by 43 countries, which agreed to freeze the CFCs production at the 1990 level and to gradually decrease production by 20% by 1994 and by 50% by 1999. In the United States, the production of CFCs was to cease by 1996 and that of HCFCs is scheduled to cease by 2020. The European Community, after amendments made in London in

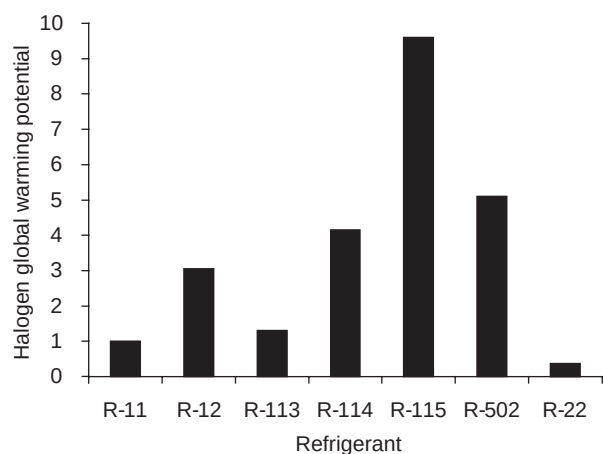


FIGURE 10 The relative halogen global warming potential of some of the most used CFCs (normalized with respect to R-11).

1990 and introduced in the Copenhagen Protocol in 1992, decided to eliminate the use of CFCs by 1995. The most promising substitution agents seem to be the HCFCs and the HFCs. Unfortunately, by increasing the amount of hydrogen in the molecule, the flammability increases, whereas by increasing the amount of fluorine, the HGWP increases. Therefore, the Copenhagen Protocol states that the HCFCs should play only a transitory role by 2030. After this deadline, only HFCs will be accepted.

The refrigerants to be replaced are R-10–R-12 and R-113–R-115. Proposed substitution agents (all of them HFCs and HCFCs) are R-141b, R-123, R-22, R-125, R-143, and R-124:

R-123 (CHCl_2CF_3) replaces R-11 and R-113: ODP = 0.5 and HGWP = 0.1.

R-141b ($\text{CH}_3\text{CCl}_2\text{F}$) replaces R-11 and R-113: ODP = 0.08 and HGWP = 0.092.

R-134a (CH_2FCF_3) replaces R-12: ODP = 0 and HGWP = 0.285.

R-290 (propane) replaces R-12: ODP = 0 and HGWP = 0.

An overview of the substitution solutions that are considered is presented in Table II.

TABLE II

Substitution Substances for the CFCs^a

Refrigerants to be replaced	Transition refrigerants			Chlorine-free alternatives for medium and long term			
	Molecular HCFCs	HCFCs mixtures		Molecular HFCs	HCFCs mixtures		Halogene-free refrigerants
		Type	Composition		Type	Composition	
R-12	R-22	R-401A	R-22/152a/124	R-134a	—	R-134a/152a	Propane (R-290)
		R-401B		R-152a			Isobutane (R-600)
		R-402A		R-32			
		R-402B		R-22/125/290			
R-508	R-22	R-403A	R-22/218/290	R-125	AZ50	R-143a/125/132	R-290, NH ₃
					FX40	R-143a/125/132/134a	
					RX4	R-32/125/134a	
					R-403B	R-143a	
R-22	R-22	—	—	R-134a	AC9000	R-32/125/134a	R-290, NH ₃
				R-32	AZ20	R-32/125	

^a Abbreviations used: CFCs, chlorofluorocarbons; HCFCs, hydrochlorofluorocarbons; HFC, hydrofluorocarbons.

5.3 Variations of the Vapor–Compression Refrigeration System

Some solutions will improve the performance or the flexibility of the vapor–compression refrigeration systems. The most used are presented here.

5.3.1 Multiple Evaporators

These systems operate at more than one evaporator temperature. The stream of liquid refrigerant coming from the condenser is divided and each fraction is throttled at a different pressure, which determines the evaporation temperature. The higher pressure fractions, after turning into vapor in the corresponding evaporators, pass through valves in which they expand in order to match the lowest pressure and to allow mixing. Such systems are used to cool different temperature loads.

5.3.2 Multistage Compression

The reason why such systems are used is to minimize the compressor work input by means of intercooling (gas cooling) between the compression stages. Vapor coming from the evaporator is compressed adiabatically by the first-stage compressor (Fig. 11). The heated vapor that results from the compression process mixes in the gas cooler with the relatively cool vapor coming from the condenser at $p_{intermediate}$. The mixture is compressed in the second stage up to the condenser

pressure, p_{high} . After rejecting heat, the vapor stream (whose temperature is lower than the temperature after the first compression stage) is divided into two streams: the main one, which goes to the main expansion valve and to the evaporator, and the secondary one, which is directed to the gas cooler. Since its pressure p_{high} exceeds $p_{intermediate}$, a throttling valve is necessary to reduce it to $p_{intermediate}$ (Fig. 11).

5.3.3 Cascaded Systems

In multistage systems, the pressure ratio p_{high}/p_{low} cannot be increased above a certain limit, which is imposed by the refrigerant's triple and critical points (p_{low} must be higher than p_{triple} and p_{high} must be less than $p_{critical}$). These limits impose a pressure range that is different for each refrigerant. Moreover, an evaporator pressure below 0.1 bar is not recommended due to the bulky compressor cylinders required for the high specific volume of the refrigerant at such a low pressure and to the special safety measures that have to be considered in order to prevent air from leaking into the installation. Therefore, for a three-stage vapor–compression refrigeration system, the highest acceptable value for the overall pressure ratio is 100–110. The temperature gap corresponding to this value is approximately 100–120°C; consequently, the minimum refrigeration temperature to be achieved is approximately

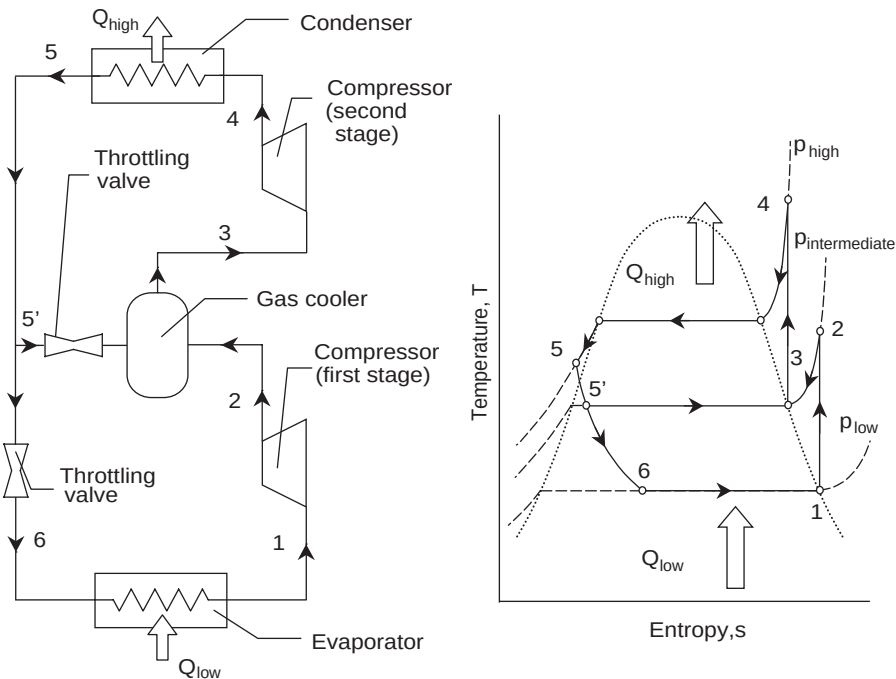


FIGURE 11 Scheme and cycle representation of a two-stage vapor-compression refrigeration system.

-70 to -80°C , considering that the condensation temperature varies from 30 to 40°C . When a single system has reached its limits, to achieve a more effective refrigeration effect, a chain of refrigeration systems using different refrigerants is necessary. Such a configuration is referred to as a cascaded system. Each of the components of a cascaded system may be single stage or multistage. With such configurations, temperatures of approximately -200°C are achievable. For instance, by cascading four systems—ammonia for the range 30 to -33°C , followed by ethylene (-28 to -104°C), methane (-99 to -160°C), and nitrogen—a temperature of -196°C can be obtained. The components in a cascade system are linked by heat exchangers, which operate as an evaporator for the “upper” component and as a condenser for the following system in the sequence.

6. ABSORPTION REFRIGERATION SYSTEMS

The idea of using an absorption fluid as a refrigerant carrier derived from the drawback of VCR systems that the gas compression requires a high work input. A pump that requires practically no work to increase the pressure in the refrigeration system replaces the complicated and work-consuming compressor.

There are two major advantages of absorption refrigeration systems (ARSs) compared with VCRs: No CFCs or HCFCs are used as refrigerants, and they use heat from different sources, such as combustion, industrial processes, waste heat (an economical solution for recovery), or solar heat.

6.1 Principles of Absorption Refrigeration

The cost of the energy supplied to the compressor represents the cost of operating a VCR system. This is why a great deal of effort has been dedicated to designing work-reducing techniques. Whatever the work-saving methods involved, as long as a gas must be compressed, a large amount of input work is required. On the other hand, increasing a liquid's pressure involves practically no work since liquids are almost incompressible. The idea of compressing a liquid in order to save work led to the development of absorption cycles. The price to be paid results from the fact that in order to accomplish the cycle, the refrigerant has to be absorbed by the liquid before compression and must be released after. The absorption process at constant pressure involves a

heat exchange: When the liquid (absorbent) absorbs the refrigerant, it releases heat (absorption heat) that has to be removed because the increase in temperature results in a decrease in the solubility of the refrigerant. When the solution containing the absorbed refrigerant must liberate it, heat input from an exterior source is necessary. Therefore, heat represents the primary energy used by ARSs.

The solution poor in refrigerant is called weak solution, whereas the solution rich in refrigerant is referred to as strong solution. An ARS is composed of a generator, in which heat is added to the strong solution to release refrigerant vapor; a condenser, throttling valve, and evaporator, which perform the same functions as for VCR systems; a pump that increases the pressure of the solution; and an absorber, in which the weak solution coming from the generator absorbs the refrigerant vapor by rejecting the absorption heat to a cooling medium (usually water). Fig. 12A shows the scheme of an ARS. There are two paths:

The refrigerant path: The high-pressure vapor leaves the generator (state 1), condenses in the condenser by rejecting heat to the surroundings, and leaves it in state 2. Its pressure is reduced in the first throttling valve (state 3) and it evaporates at low pressure in the evaporator by removing heat from the cooled space. State 4 corresponds to saturated low-pressure vapor. The vapor is absorbed in the absorber by the weak solution.

The solution path: The pressure of the strong solution (state a) is increased by the pump (state b). The absorption occurs in the absorber and is facilitated by the heat removal performed by the water that flows through the cooling coil. The strong solution enters the generator, where heat is added in order to perform the separation process of the vapor from the solution. The high-pressure vapor leaves the generator, and the weak solution that results from the separation process (state c) is throttled in the second throttling valve to the absorber pressure (state d).

To save energy, a regenerative heat exchanger (also called economizer) may be placed on the solution path (Fig. 12B). The weak solution that leaves the generator still contains enough thermal energy to preheat the weak solution entering the generator. Moreover, the process of solution enrichment occurring in the absorber generates heat that must be removed. Precooling of the weak solution can improve this process. The strong solution

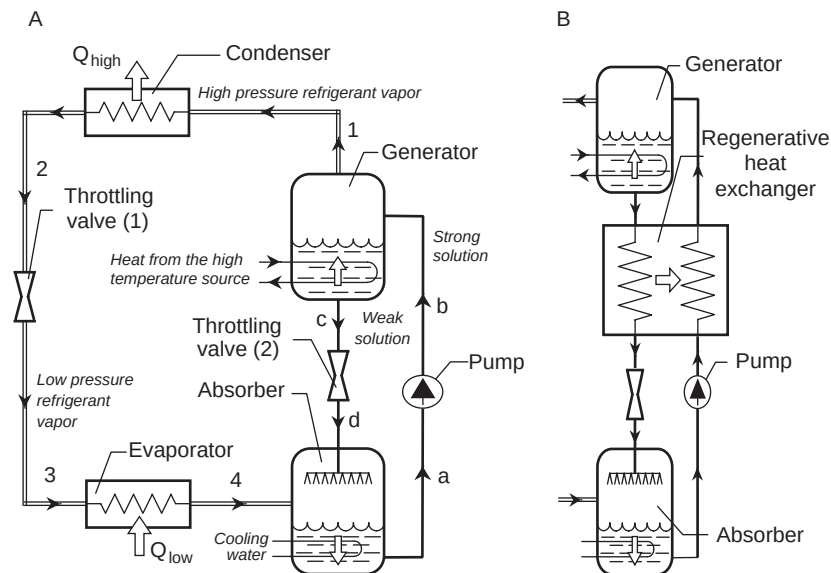


FIGURE 12 Scheme of an absorption refrigeration system: (A) without heat regeneration; (B) with heat regeneration (only the solution path is schematized).

preheating before the generator saves energy by “returning” heat to the separation process that occurs in the generator.

The COP of the ARS is

$$\text{COP} = \frac{Q_{low}}{Q_{high} + W_p},$$

where Q_{low} is the heat removed from the low-temperature body, Q_{high} is the heat rejected to the surroundings (the high-temperature body), and W_p is the work input of the solution pump.

Supposing that the heat transfer processes are isothermal and reversible and that the pump work input is negligible, the previous expression of the COP becomes

$$\text{COP} = \frac{T_{high} - T_A}{T_{high}} \times \frac{T_{low}}{T_A - T_{low}},$$

where T_{low} and T_{high} are the temperatures of the cooled space, respectively, of the surroundings, and T_A is the temperature of the heat removal in the absorber. Note that this equation represents the upper limit of the COP (under ideal conditions). Due to the irreversibilities, the actual COP is less than this maximum value. In terms of actual COP, ARSs are generally poor compared with VCRs. For instance, the COP of a lithium bromide–water system is approximately 0.6, which can be improved by different techniques but with increased capital costs.

Most ARSs use the pairs water–ammonia and lithium bromide (LiBr)–water. They can be coupled

TABLE III

Absorbent–Refrigerant Pairs Used in Absorption Refrigeration Systems

Absorption Medium	Refrigerant
Water	Ammonia
Water	Methyl-amine
Lithium chloride or bromide	Water
Paraffin oil	Toluene or Pentane
Sulfuric acid	Water
Dimethyl ether or tetraethylene glycol	Methylene chloride
Sodium or potassium hydrate	Water

with vapor–compression refrigeration systems and are used especially in the chemical industry and in air-conditioning. Their main advantage is the simultaneous refrigeration and warm water supply. The most common pairs used in ARSs are presented in Table III.

6.2 Water–Ammonia System

This system is used especially for small refrigerators with gas or electric heating. There are three major differences with respect to the LiBr–water systems:

1. The water–ammonia systems operate at high pressures (4–5 bar), whereas the LiBr–water machines operate at 0.04–0.08 bar.

2. Whereas LiBr is not volatile, water releases vapor during the distillation process, and therefore a rectifier is required before the condenser to separate water vapor from ammonia vapor.

3. The cooling medium for the water–ammonia system is air, whereas for the LiBr–water machines water is used as coolant.

6.2.1 Platten–Munters (Serivel) System

This is a water–ammonia ARS used especially for household refrigerators and has no moving parts (pumps). The main feature of this system is the presence of an inert gas (hydrogen) that has two main functions: (i) It plays the role of an equalizing medium for the pressures in the system and thus it eliminates the necessity of a pump between the absorber and the generator, and (ii) it acts as a carrier for the refrigerant.

The pressure is the same in the entire system, namely the condenser pressure. Ammonia has different partial pressures in different regions of the machine. The ammonia vapor that leaves the generator and the rectifier chases the hydrogen from the condenser. After the condenser, liquid ammonia directly enters the evaporator, without any throttling. Here, ammonia evaporates and its vapor diffuses among the hydrogen molecules. The evaporation occurs due to the difference between the partial pressure of the refrigerant vapor in the close vicinity of the free surface of the liquid and the partial pressure of the refrigerant vapor in the hydrogen–ammonia mixture. This mixture is heavier than pure hydrogen. It moves downward and leaves the evaporator entering the absorber, where the weak solution absorbs the ammonia vapor from the mixture, which becomes lighter and returns to the evaporator. The strong solution leaving the absorber passes through a coil, in which a fraction of it releases vapor that “pumps” the solution to the generator by the thermosyphon effect. The heating of the generator is performed either electrically or with combustion gases. The separation process occurs in the generator and the weak solution returns to the absorber by gravity. The ammonia vapor enters the condenser and the cycle is repeated.

6.3 Lithium Bromide–Water System

The refrigeration systems belonging to this family can be classified from the standpoint of the origin of the heat source in indirect-fired systems, which use steam or hot liquid; direct-fired systems with heat

from combustion gases; and heat-recovery systems, which use waste heat. Alternative heat sources, such as solar or geothermal energy, are also used.

The absorbent (LiBr) acts as a refrigerant carrier from the low-pressure region to the high-pressure region of the machine. For water chilling, the COP of a LiBr–water ARS varies in the range 0.65–0.7. Heat input temperatures for single-stage machines are usually 95°C, whereas for two-stage ones they are approximately 150°C.

ARSs that use this pair usually serve as water chillers for air-conditioning applications. Water is chilled to approximately 5°C. The condenser temperature is 40–45°C, whereas the evaporator temperature is 4–7°C. Temperatures below the water’s triple point are not reachable since water (the refrigerant) freezes.

Low-power units used in unitary air conditioners are natural gas fired. Very high-power units with a refrigeration capacity of up to 6 MW use 1.5–2 bar of steam as a heat source and are cooled with water at 26–30°C.

6.3.1 Carrier Absorption Refrigeration System

ARSs operate at two pressure levels, p_{high} and p_{low} . The pressure is p_{high} in the generator and condenser and p_{low} in the evaporator and absorber. Therefore, the components operating at the same pressure level may be grouped in the same vessel. The Carrier system uses this arrangement, as shown in Fig. 13.

The LiBr–water solution boils in the generator by removing heat from the high-temperature heat source (steam, hot water, solar heat, or waste heat). Only water vapor is released. The vapor enters the condenser, where it condenses by rejecting the latent heat of evaporation to the cooling water that flows through a coil. The resulting liquid refrigerant passes through a throttling valve, where its pressure is reduced and it turns into a liquid–vapor mixture, which enters the evaporator. Here, the mixture evaporates by removing heat from the cooled agent (chilled water). To improve the process, the liquid fraction of refrigerant is recirculated by a refrigerant pump and sprayed onto the heat transfer surface of the cooling coil. The vapor passes to the absorber, where it is put into contact with the sprayed weak solution and is absorbed. The strong solution that results is pumped and flows through the economizer, where it removes heat from the downstream flowing weak solution and enters the generator. The process is then repeated.

The coolant (water) first flows through the absorber, where it removes the absorption heat from

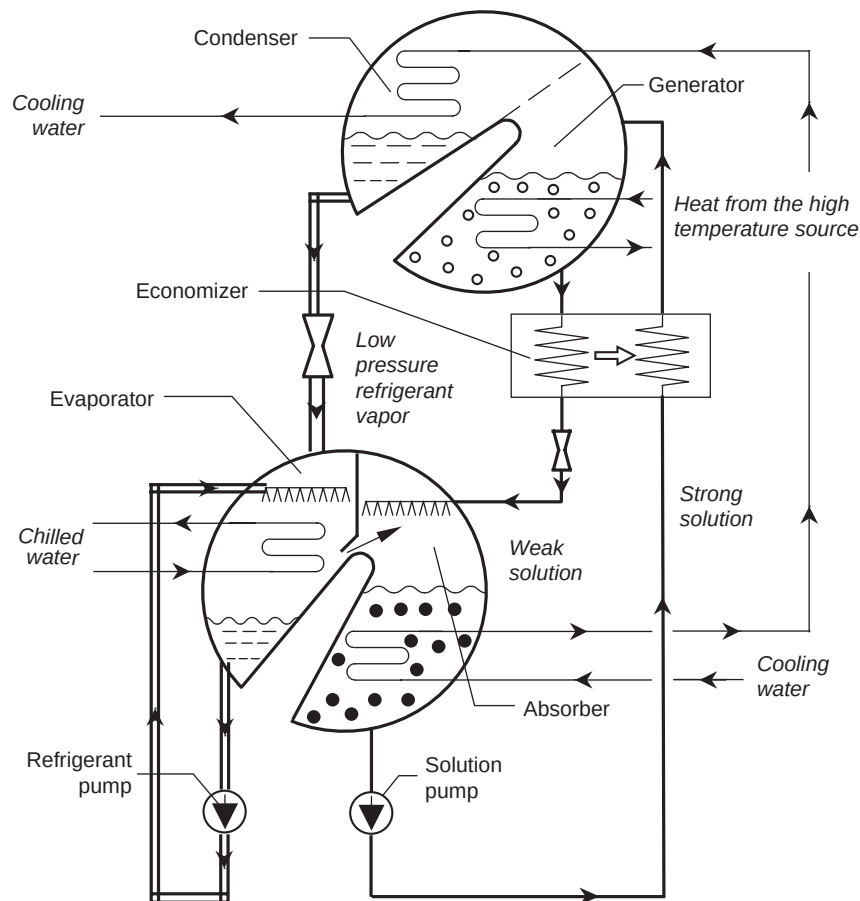


FIGURE 13 Scheme of the Carrier absorption refrigeration system.

the solution, and subsequently goes to the condenser, where it is heated by the condensing refrigerant.

In contrast to the Carrier system, which uses two vessels, in the Trane system a single vessel contains all the components of the absorption refrigeration machine. A comparison between the water-ammonia and the LiBr-water systems is provided in Table IV.

7. STEAM JET REFRIGERATION

The specific feature of the steam jet refrigeration (SJR) is the use of water as refrigerant. From the thermodynamic standpoint, water is a very good refrigerant since it has a high latent heat of evaporation and it is cheap, nontoxic, nonflammable, and available in unlimited quantities. On the other hand, water exhibits two important shortcomings: very low saturation pressures and very high specific volumes at usual temperatures. For these reasons, the use of water as refrigerant is limited to temperatures in the range 8–11°C and very rarely

below 4°C. To obtain an evaporator temperature of 7°C, p_{low} must be approximately 1 kPa. Consequently, very large specific volumes (on the order of 129 m³/kg) are involved in the compression process, and only centrifugal compressors are able to manipulate them. Another option is to use steam ejectors, which represent an economically convenient solution when steam is available and is not produced just for this purpose.

SJR is actually VCR with the mechanical compressor replaced by an ejector (or jet compressor). Although much less efficient than centrifugal compressors, jet compressors are simple (no moving parts) and inexpensive. The condenser operates at subatmospheric pressure and air leakage into the condenser is to be expected. For this reason, a secondary ejector driven by a fraction of the steam main stream removes the noncondensable gases. The high-pressure steam used to compress the low-pressure steam coming from the evaporator must also condense; therefore, the heat transfer surface area of the condenser must be large enough to

TABLE IV
Comparison between the Water–Ammonia and Lithium Bromide–Water systems

Criterion	Water–ammonia	Lithium bromide–water
Saturation temperatures difference	Relatively low	Very large
Refrigerant vapor purity	Poor The vapor released in the generator is a mixture of water and ammonia vapor and therefore a rectifier is required.	Very high The vapor in the generator is pure water vapor. No rectifier is required.
Absorption rate	High No recirculation is needed.	Low LiBr slowly absorbs water and therefore a recirculation of the weak solution in the absorber is necessary.
Toxicity	High	Relatively low
Corrosion	Noncorrosive toward ferrous metals Ammonia corrodes copper and its alloys.	Corrodes nonferrous metals
Heat supply shutoff	Accepted The cooling of the solution due to the lack of heat input does not result in crystallization, which occurs at temperatures far below the ambient temperature.	Not accepted The strong solution crystallizes at ambient temperature. In such situations, the installation has to be emptied and the solution is stored in heated reservoirs.

achieve this task. As a consequence, the size of the condenser must be three or four times the size of a VCR condenser.

The capacities of SJR systems are in the range 35–3500 kW. Typical applications of such systems are foods drying, concentration of fruit juices, and dehydration of chemicals. Direct vaporization is used in order to cool the substance; therefore, no heat exchanger or indirect cooling process are required, making SJR more suitable than VCR.

SJR systems are simple, require low capital costs and a low operating cost if using waste steam, reliable (safe) due to the lack of moving parts, and use an inexpensive refrigerant (water). On the other hand, they require a high consumption of steam and cooling water for the condenser, are difficult to operate under conditions of variable load, and have a limited area of applications compared with the other refrigeration systems. Their only worthwhile application is cooling by evaporation in drying processes.

A SJR system is presented in Fig. 14. The liquid water entering the evaporator evaporates at constant pressure by removing heat from the cooled body. The vacuum generated in the ejector by the high-pressure steam sucks the resulting vapor. The compression effect is accomplished in the jet compressor's diffuser section by the conversion of the kinetic energy of the jet into potential energy of pressure. The high-

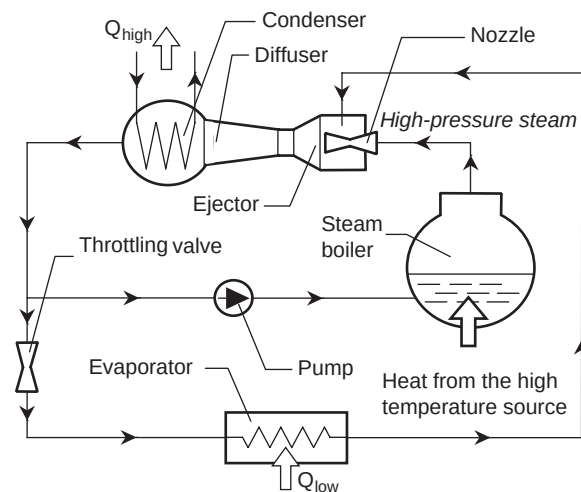


FIGURE 14 Steam jet refrigeration system.

pressure vapor condenses in the condenser and the resulting liquid is divided into two streams: One fraction is pumped into the steam boiler, where it turns into steam, and the other passes through the throttling valve and evaporates in the evaporator by removing heat.

Adjustment of the refrigeration power can be performed by varying the number of ejectors working at the same time (gross tuning) and by means of the steam flow rate per ejector (fine-tuning).

8. OTHER METHODS OF REFRIGERATION

Other refrigeration techniques are used, but only within a narrow area of applications: electronic components cooling, air liquefaction, refrigeration of small spaces, and aerospace.

8.1 Reversed Stirling Cycle (Philips Machine)

The power Stirling cycle was patented in 1816. The reversed cycle was applied in refrigeration after World War II by the Philips Company. Basically, the cycle is composed of two isothermal and two isometric (constant-volume) processes.

An interesting attribute of the Stirling cycle is that it occurs in a closed system: The working fluid is contained in the machine and there is no exchange of substance between the exterior and the refrigeration system. Another beneficial consequence is the possibility to choose the optimal fluid (from the thermodynamic standpoint) since cost is of no importance in this case due to the small quantity of refrigerant required. Moreover, because under ideal conditions all heat transfer processes between the system and the heat source and sink occur isothermally, the cycle's efficiency equals the Carnot efficiency. Obviously, under actual conditions, the Stirling's cycle efficiency is less than the ideal one due to the irreversibilities that inherently occur in every process.

The main feature of the cycle is heat regeneration, namely the internal flow of heat, which is first stored in a structure called a regenerator and is discharged later to the gas circulating through this unit. The regenerator consists of a heat-storage material (ceramic mesh, wire, or metallic gauze) through which the gas is forced back and forth during the isochoric processes. The gas rejects heat to storage medium (cooling) or removes heat from it (heating). In this manner, heat is "lent" to the storage mass of the regenerator during the isochoric gas cooling and is taken back later from it during the isochoric gas heating.

The machine (Fig. 15) is of the reciprocating type and consists of a cylinder with two pistons (A and B) and a regenerator (R). By means of a sophisticated mechanical system, the two pistons perform inter-related motions that allow the processes that comprise the cycle to occur.

The principle of the cycle operation is schematized in Fig. 15 in correspondence with the T - s plot of the processes. The first process is the isothermal compression (1-2), which occurs as piston B is moved to the left while the coolant removes the heat that is generated as a consequence of the compression work and thus the temperature T_{high} of the gas remains constant. During this time, piston A is held in position. The following process (2-3) is performed by moving both pistons at the same rate to the left. The volume of gas remains constant and the evolution is isochoric (isometric). The gas is forced

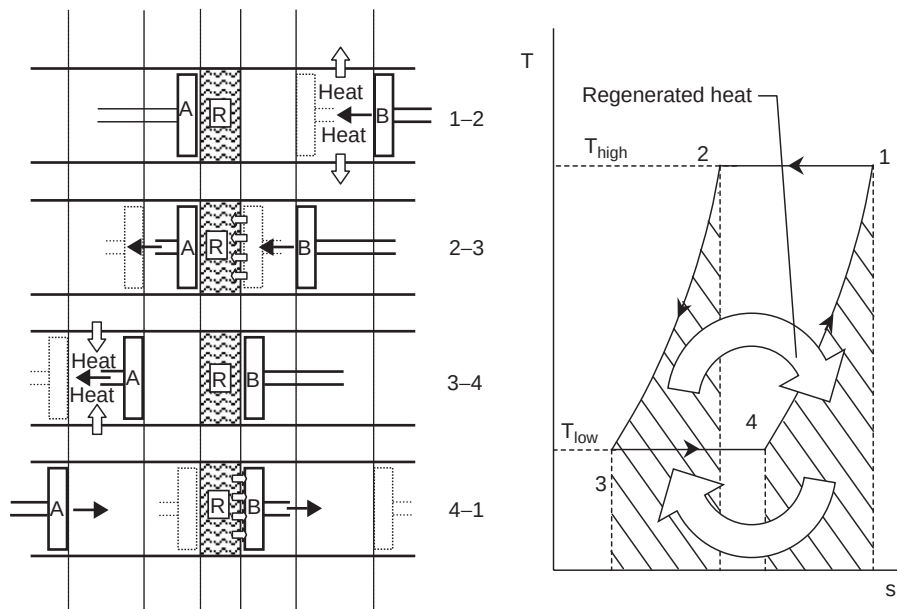


FIGURE 15 Schematic representation of the reversed Stirling cycle operation.

through the regenerator and rejects heat to the storage medium, which is colder as a result of the previous passage of gas. Isothermal expansion (3–4) follows, which is performed by blocking piston B and moving piston A to the right while heat is removed from the cooled space at $T_{\text{low}} = \text{constant}$. The last process of the cycle is isochoric gas heating (4–1). Both pistons move at the same rate to the right and the volume of gas is constant. By passing through the regenerator R, the cold gas removes the heat that has been stored during the previous isometric process and regains the temperature from the beginning of the cycle (T_{low}).

The Philips machine is used for temperatures in the range 223–13 K (-50 to -260°C). Most of its applications are in cryogenics. The actual machine, designed by Köhler and Jonkers, differs from the previously discussed scheme from the standpoint of the pistons' movement, which is continuous (the pistons of the ideal machine move intermittently). Helium and hydrogen are the usual working fluids. For temperatures higher than -130°C , air may also be used. The most usual applications are in the area of scientific research, gas liquefying, etc. The Philips refrigeration machine is characterized by very low losses and therefore its efficiency in the temperature range -150 to -90°C is approximately 0.4. A machine with a relatively small size (cylinder inner diameter = 70 mm and piston stroke = 50 mm) can liquefy approximately 5 kg of air per hour, with the necessary power input being approximately 6 kW.

8.2 Open-Cycle Systems

8.2.1 Gas Expansion in a Field of Centrifugal Forces (Ranque Effect)

The French physicist Ranque observed this phenomenon for the first time in 1933, which is why it is referred to as the Ranque effect. The vortex tube device of a cylindrical tube is equipped with a tangential nozzle and a cone, as in Fig. 16. As the gas expands in the nozzle, a very intense circular current develops, characterized by a strong radial velocity gradient. The compression of the gas close to the tube wall heats the gas, whereas the circulation

becomes longitudinal near the tube's axis. As a result, a separation of streams occurs: The hot one leaves the tube at the right end and the cold one is collected at the exit orifice from the left end. The axial position of the cone controls the flow rate and thus the temperatures of the two streams. Turbulence and viscosity are the source of the phenomenon: The kinetic energy of the gas particles has important variations, and thus the gas is compressed and expanded adiabatically, performing a refrigeration cycle that pumps heat to the peripheral regions of the gas by cooling the regions situated on the tube axis (Fig. 16).

The parameters that influence the performance of the vortex device are the tube's geometry (length and diameter), the diameter of the cold gas escape orifice, the nozzle's section area and type, the temperature and pressure of the compressed gas, and its physical properties. For instance, for an air pressure of 11 bar at an inlet volumetric flow rate of $30.5 \text{ m}^3/\text{h}$, by using a tube with a certain geometry one can obtain a fraction equal to 0.2 from the total flow rate at -41°C .

8.2.2 Reversed Thermoelectric (Peltier) Effect

The direct thermoelectric effect was discovered in 1821 by the German physicist Thomas J. Seebeck. It consists of the apparition of an electromotive force as the junctions of two wires made of different electrical conductors (the ensemble is called thermocouple) are placed in mediums with different temperatures.

The reversed thermoelectric effect was discovered in 1834 by the French physicist Jean-Charles Anastase Peltier. By connecting a thermocouple to a direct current (DC) source, one junction heats and the other cools. By inverting the current's direction, the cold and hot junctions switch functions.

The apparition of a temperature difference between two junctions involving different metallic conductors when connected to a DC source is referred to as the Peltier effect. The explanation resides in the fact that different electrical conductors exhibit different energy levels of their conduction electrons. When forced to pass from the high-energy region to the low-energy region under the influence of an electric field, the electrons reject their excess energy as heat. This phenomenon occurs at the boundary between the two conductors, which are put into contact by welding. The opposite process takes place at the other junction: Low-energy electrons absorb from the surroundings the necessary energy to pass to the high-energy region, and this phenomenon manifests as heat removal from the environment.

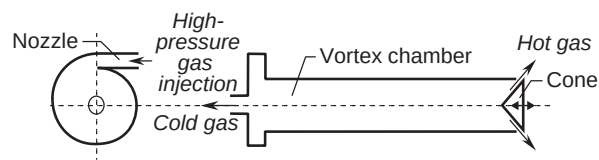


FIGURE 16 The vortex (Ranque) tube.

The cooling effect is poor in conductors, whereas in semiconductors its intensity is important. The maximum cooling efficiency is reached when one of the semiconductors is of the n type and the other is of the p type. Typical p-type semiconductors used for this purpose are alloys of Sb_2Te_3 and Bi_2Te_3 , and n-type semiconductors are alloys of Bi_2Te_3 and Bi_2Se_3 . With such couples, temperature differences of up to 60°C have been obtained. By connecting more thermoelectric couples in series (battery arrangement: the cold end of one of them cools the hot end of the other), the temperature gap can be increased.

Thermoelectric refrigeration batteries are used in low-power as well as high-power applications. For small systems, optimization aims to get the highest power possible by sacrificing efficiency. High-power systems are designed to achieve the highest performance in terms of efficiency. Other potential applications of thermoelectric refrigeration systems are small air-conditioning systems, household refrigerators, and aerospace.

The cooling effect is reduced by heat generation at the passage of the electric current through a conducting body (the Joule–Lenz effect) and by the conduction heat transfer between the two junctions.

The advantages of thermoelectric refrigeration systems result from the fact that they require no refrigerant, have no moving parts, are simple, the refrigeration power can be as small as necessary, the refrigeration element has a very small size (applications in surgery, electronics, and dentistry), and the refrigeration temperature can be easily controlled by varying the current's intensity. On the other hand, they have poor shock resistance and are expensive.

9. CRYOGENICS

The term cryogenics derives from the Greek *kryos*, for icy cold. The upper limit that characterizes the cryogenic region of the temperature scale is subjective. If the delimitation criterion takes into consideration practical reasons, then this upper limit can be established at a higher than usually declared value (120 K) in order to include natural gas liquefaction. The cryogenics region of the temperature scale 0–120 K can be divided into subregions according to the criteria of the cryogenic method and refrigerant as follows:

- 63–120 K: The industrial cryogenics domain. The cryogenic methods are based on thermodynamic

state-change processes of the usual refrigeration type, as described previously.

- 1–60 K: Thermodynamic methods combined with additional vacuum processes that characterize the domain of liquefaction of hydrogen, helium, neon, nitrogen; also of industrial importance.

- 0.3–1 K: The domain of laboratory cryogenics.

- <1 K: The cryogenics domain of advanced scientific research. The methods used are of the magnetic type. Temperatures on the order of 10^{-8} K have been obtained. The third law of thermodynamics (Nernst, 1906) states that absolute zero cannot be reached whatever method used.

Cryogenics is most important in industrial applications: liquefaction, gas separation, and superconductivity. Cryogenics is also on the cutting edge of scientific research: nuclear and elementary particles physics, physics of the solid, cryobiology, and cryosurgery.

9.1 Cryogenic Cycles and Methods

Some of the previously described refrigeration techniques are also applicable to the cryogenic temperatures' domain. In gas-expansion processes, the gas, after being compressed and rejecting heat, expands in an expansion machine. Compared with throttling, this method has a higher cooling intensity, expressed as the temperature drop per pressure drop (degrees/bar). In the cryogenic cycles without the refrigerant's phase change, the latter remains gaseous as it undergoes the processes that comprise the cycle. Usually, these cycles correspond to low- and medium-power refrigeration systems (also known as gas refrigeration machines) and are used in laboratory applications. The most used is the Philips machine, whose cycle (the reversed Stirling cycle) was previously described. By using helium as refrigerant, such machines can reach 77 K in a single stage and 12 K in two stages.

In the following sections, specific cryogenic methods are presented. Some of the most used cryogenic agents are shown in Table V along with some important physical properties.

9.1.1 Joule–Thomson Effect

By being throttled, the gas undergoes a pressure reduction and under certain conditions its temperature decreases. The phenomenon is called the Joule–Thomson effect and is accompanied by a temperature change of the gas. If the process occurs within a temperature and pressure range that ensures that the

TABLE V
Main Cryogenic Agents and Important Physical Properties

Cryogen	Normal boiling point (K)	Critical temperature (K)	Critical pressure (kPa)	Triple point temperature (K)	Triple point pressure (kPa)
Helium	4.22	5.28	227		
Hydrogen	20.39	33.28	1296	14.0	7.2
Nitrogen	77.33	126.17	3385	63.22	12.55
Carbon monoxide	82.11	132.9	3502	68.11	15.38
Argon	87.28	151.2	4861	83.78	
Oxygen	90.22	154.8	5081	54.39	0.14
Methane	111.72	190.61	4619	90.67	11.65
Krypton	119.83	209.4	5488	116.0	73.22

effect is positive (the gas temperature decreases as its pressure is reduced), then throttling results in gas cooling and can be successfully used to liquefy gases. The following requirements must be met:

1. The gases involved must have very low saturation temperatures and pressures.
2. In order to bring the gas into a state characterized by a positive Joule–Thomson effect, a foregoing preparation is necessary: the gas has to be precooled in auxiliary classical refrigeration systems.
3. To obtain very low temperatures, very high pressure drops are necessary.

For instance, for hydrogen, the pressure value before throttling must equal 110 bars after a precooling below 190 K (approximately -83°C) to obtain a positive Joule–Thomson effect.

9.1.2 Linde Method

Basically, the configuration of the cryogenic system is that of a classical VCR machine. The specific features of the cryogenic system are the aftercooler that replaces the condenser and the counterflow heat exchanger placed between the aftercooler and the throttling valve. In the aftercooler, the gas is brought to the ambient temperature. The function of the counterflow heat exchanger is to perform a supplementary cooling of the high-pressure gas that leaves the aftercooler. The coolant is cold gas collected from the separation vessel placed after the throttling valve. A low-quality liquid vapor mixture results after the throttling process. The mixture enters a separation vessel, where the liquid phase is gravitationally collected at the bottom and withdrawn, whereas the vapor occupies the space above the liquid. The vapor (which is at the saturation temperature) is circulated

through the counterflow heat exchanger and removes heat from the stream of high-pressure gas coming from the aftercooler. This supplementary cooling significantly increases the liquefaction efficiency.

Adding a second throttling and a precooling of the compressed gas after the compressor can improve the basic configuration described previously. Such improvements reduce the necessary power input of the cryogenic system.

9.1.3 Claude Method

This technique uses an expander and is an improved variation of the Linde method. By expanding a gas in a machine, the work output combined with the pressure decrease leads to a stronger cooling than that of the simple pressure destruction by throttling. Consequently, after being cooled in the aftercooler, the gas undergoes a primary cooling in a counterflow heat exchanger by rejecting heat to a stream of cold gas coming from the second-stage counterflow heat exchanger. After this primary cooling, the gas stream is divided: A fraction enters the second-stage heat exchanger, whereas the remainder expands in the expander, resulting in a colder gas, which is returned to the main coolant flow path that enters the second-stage heat exchanger. The vapor collected at the upper end of the separation vessel is directed to the final counterflow heat exchanger, where it cools the gas stream that leaves the second-stage heat exchanger. By cascading several Linde and Claude cryogenic systems operating with different refrigerants, one can obtain liquefaction of helium and an additional cooling in helium down to 0.7 K.

9.1.4 Diffusion Pump Method

This device is used to liquefy helium and is based on the active coal's property of adsorbing or desorbing

helium. Active coal has a very porous structure, which allows it to adsorb gases in large amounts. The quantity of gas adsorbed increases as temperature decreases and as pressure increases. As a consequence, a gas is adsorbed at low temperature. By heating the adsorbent, the gas is released. This property allows the adsorbers to operate in adsorption–desorption cycles. The adsorption process is exothermic (i.e., it releases heat), whereas the desorption is endothermic, requiring heat in order to occur. These phenomena are used to liquefy helium in cryogenic systems in which the compressor has been replaced by a diffusion pump.

The active coal is saturated with helium at normal or higher pressure and the heat of adsorption is evacuated by cooling. The system (coal + adsorbed helium) is then thermally insulated and undergoes a process of desorption by creating vacuum in the space that confines it. As a result, the temperature decreases, and the refrigeration effect occurs. Using such a technique, a temperature of 2 K has been obtained in liquid helium.

9.1.5 Thermomagnetic Effect

A substance that is not a magnet in the absence of a magnetic field is called paramagnetic. Only few substances are paramagnetic: the crystalline hydrates of some rare minerals (e.g., $\text{Gd}_2\text{SO}_4 \cdot 8\text{H}_2\text{O}$) or the salts of the metals from the platinum or aluminum groups.

When subjected to an exterior magnetic field, the microscopic dipole magnetic moments of the paramagnetic material oppose this action and consequently the magnetic field performs work to change their orientation. At the macroscopic level, this phenomenon manifests as heating of the substance. If one removes this heat with a coolant, the magnetization process will be isothermal. Switching off the magnetic field, the substance undergoes a demagnetization process and the energy stored in the substance as a consequence of its opposition to the action of the field is transferred to the magnetic dipoles to regain their chaotic orientation. If the substance is thermally insulated, the result is an intense cooling. The phenomenon is called the thermomagnetic effect and the method is referred to as adiabatic demagnetization of paramagnetic substances.

A second phenomenon occurs during the magnetization–demagnetization cycle of the paramagnetic body. The internal energy of a substance depends not only on the kinetic and potential energies of the molecules but also on the magnetization. As magne-

tization increases, more order is transferred to the system by orienting the elemental magnets along the magnetic field lines and thus the chaotic movements of the molecules are limited. The number of degrees of liberty of the thermal motion is correspondingly reduced. By this action, an internal magnetic field develops, called the internal Lorentz field. The internal energy of the body decreases as a consequence of magnetization. During the adiabatic demagnetization phase, not only does the magnetization work have to be redistributed to the molecules to regain their initial chaotic orientation but also the amount of internal energy corresponding to the degrees of liberty lost during the magnetization process has to be redistributed. An additional temperature decrease occurs due to this internal energy redistribution process.

By using the adiabatic demagnetization of paramagnetic salts, temperatures in the range $0.3\text{--}10^{-3}\text{ K}$ can be obtained and maintained for a short time.

To obtain temperatures less than 10^{-3} K , the adiabatic demagnetization of the nuclear state of a paramagnetic salt has been used. The phenomenon is basically the same as in the case of the atomic paramagnetism. Using adiabatic demagnetization of the nuclear spin system of copper, a temperature of 10^{-8} K was attained in the 1980s.

9.2 Applications

9.2.1 Air Separation

Air components such as oxygen, nitrogen, and argon are used in different applications. Nitrogen is used in metallurgy (as inert atmosphere), electronics, chemical processes, and deep refrigeration.

Oxygen is used in metallurgy, welding, chemical industry, and medical applications. Argon is used as inert atmosphere in electric bulbs and welding.

Dry air is a mixture of three major gases (in moles, 78.03% nitrogen, 20.99% oxygen, and 0.934% argon) and other components, such as carbon dioxide, rare gases, hydrogen, and helium (all less than 0.05%). Air liquefaction constituted the first application of the Linde method in 1895. Before being processed, air is dried up and carbon dioxide is removed since water vapor and CO_2 freeze during air liquefaction, endangering the operation of the air-separation unit. The separation of the three main components is based on the distillation process. A two-staged distillation column operating at two pressure levels separates oxygen and nitrogen, whereas argon is distilled in a separated column. A typical air-separation plant supplies liquid and

gaseous oxygen, which are collected at the lower end of the low-pressure stage of the column, and liquid and gaseous nitrogen, which are extracted from the medium- and high-pressure regions of the second-stage column. A turbine expander supplies the refrigeration effect necessary for the air separation. Less important refrigeration is obtained by means of the Joule–Thomson effect.

Total air-separation plants provide oxygen, nitrogen, argon, and rare gases (He, Ne, Kr, and Xe).

9.2.2 Liquefaction of Natural Gas

Two reasons justify the liquefaction of natural gas. First, the demand is not constant, exhibiting a peak in midwinter. To meet this increased demand, the distribution piping should be designed accordingly, which involves high capital costs. It is more economical to design the distribution grid according to a medium demand, to store liquefied gas in large storage tanks, and to vaporize and add it to the network when necessary. Second, the oil/gas field is usually far away from the users or in a location that does not economically or technically justify a pipeline for collecting the extracted gas (e.g., offshore extraction platforms). In such cases, gas is liquefied and transported to a facility near the user area, where it is vaporized and injected into the distribution network.

The refrigeration effect necessary for liquefaction is achieved by a variety of methods. In some cases, methane is used as a refrigerant in intermediate refrigeration.

9.2.3 Helium Liquefaction

Helium was the last of the gases to be liquefied in 1908 by Kammerlingh Onnes. Helium is used in welding, aerospace, superconductivity, inert atmosphere applications, medicine, and advanced scientific research. Helium's concentration in air is well below the economically acceptable level. However, it is present in much higher concentrations (up to 15%) in natural gases. Helium's liquefaction requires special cryogenic techniques since its saturation temperature is the lowest among all substances (Table V). Due to the fact that at room temperature helium has a negative Joule–Thomson effect (it heats when throttled), previous precooling below the inversion temperature (the temperature above which the Joule–Thomson effect changes slope) is necessary. This is performed by means of a Claude process, which brings the gas to 20 K at 3000 kPa. Liquefaction is then accomplished by means of an expansion valve.

9.2.4 Superconductivity

Helium has two stable isotopes, He_I (atomic mass = 3) and He_{II} (atomic mass = 4), which are present in different concentrations. Its normal boiling point is 4.23 K, and below this temperature it is a colorless lightweight liquid with a very low viscosity. The separation of the two isotopes can be done at temperatures lower than 2.17 K, which is the threshold below which He_I is in a normal state, whereas He_{II} is in the superfluid state. Helium is the only substance that can maintain its liquid state at near absolute zero and has no triple point. Solid helium can be obtained only under high-pressure conditions (at $T < 1.0$ K, the pressure must be higher than 2500 kPa to obtain solid He_{II} or higher than 3000 kPa for He_I). The higher the temperature, the higher the pressure: For instance, at $T \approx 60$ K, the solidification pressure is approximately 10^6 kPa.

Superfluidity is a special property of liquid He_{II} at $T < 2.17$ K. The liquid's unusual behavior consists of the fact that its flow velocity through capillary tubes depends only on the temperature and the tube diameter, and not on the pressure drop and the tube length. The flow occurs without friction, with the viscosity of the liquid becoming extremely low. The phenomenon was discovered in 1932 by the Russian physicist P. L. Kapitsa, who called it superfluidity. At 5.072 kPa and $T = 2.174$ K, helium's thermal conductivity exhibits a radical change: It suddenly becomes 3×10^6 times higher than the thermal conductivity of other liquids at the same temperature (liquid helium included) or 1000 higher than copper's conductivity at ambient temperature.

In normal electrical conductors, the resistance decreases as temperature decreases. In 1911, by studying the electrical behavior of a mercury sample in liquid helium, the Dutch physicist Kammerlingh Onnes discovered that the sample's electrical resistance disappeared suddenly. It was the first time that the phenomenon called superconductivity was noticed. Further research on this phenomenon was performed and it was found that it occurs in a wide range of metals and alloys. Practically, the resistance equals zero. This fact has been verified by immersing a metal ring in liquid helium and injecting an electric current. The current can run through the ring undiminished for months.

Superconductivity is used mainly in magnet applications (plasma physics and electrical machines). An interesting application may be the use of superconductive magnets for magnetic levitation in high-speed railroads (500 km/h).

Other potential applications are in the field of electric energy transport. Power plants are situated as close as possible to the fuel source, which is generally far away from users. Consequently, the energy has to be transported through metallic cables. Due to the losses caused by the metal's electrical resistance, economic limitations dictate that the transported power at 1.5×10^6 V be less than 10^6 kW. The use of superconductive cables may be economically acceptable for energy transport at distances on the order of 200 km for AC and 50–800 km for DC. In the case of very high-power superconductive energy transport, the economy may be approximately 5 or 6 MWh/km.

For obvious economic reasons, the research in the field focuses on different alloys and semiconductor combinations in order to develop substances that reach the superconductive state at the highest possible temperatures, preferably 100 K (liquid nitrogen).

10. AIR-CONDITIONING

The following are the most important developments in the evolution of air-conditioning and air-conditioning systems:

At the end of the 19th century:

- Founding of the scientific hygiene (Max von Pettenkofer, Germany)
- Research on air exchange, hygrometry, and air moisture
- The use of electric motors for air fans in large buildings; air filtration
- By 1890, air humidification with steam and later by water spraying; beginnings of air-conditioning.

20th century:

- By 1900, the first air-conditioning systems with preheater, humidifier, and reheater (United States); W. H. Carrier (1876–1950), father of air-conditioning
- The first temperature and humidity control systems (pneumatic and electrical)
- By 1920, the upsurge of air-conditioning; central air-conditioning systems for comfort (theaters, office buildings, and conference halls) and industry, especially for hygroscopic materials processing (tobacco, paper, and textiles)
- By 1930, the first unitary systems that include all the components; the use of freons for the refrigeration side; room air conditioners

TABLE VI

The Standard Dry Air Composition

Component	Volume fraction
Oxygen	0.2095
Nitrogen	0.7809
Argon	0.0093
Carbon dioxide	0.0003

After 1945, very rapid development of air-conditioning systems: dual-path systems, air-and-water systems with induction devices, and fan-coil units.

10.1 Basic Principles

The standard dry air composition as established in 1949 by the International Joint Committee on Psychrometric Data is presented in Table VI. The water vapor content of the air may vary from zero (dry air) to a maximum (saturated air) that depends on the temperature at a certain atmospheric pressure.

The humidity ratio W of the air is the ratio of the mass of water vapor (m_v) to the mass of dry air (m_a):

$$W = \frac{m_v}{m_a} \left[\frac{\text{kg water vapor}}{\text{kg dry air}} \right].$$

The relative humidity ϕ is the ratio of the mass of water vapor (m_v) contained in the air to the mass of water vapor at saturation (m_s) at the same temperature:

$$\phi = \frac{m_v}{m_s}.$$

The relative humidity varies in the range 0–1 (0, dry air; 1, saturated air).

The dew point temperature is the temperature below which water vapor condensation occurs if the air is cooled at constant humidity ratio. The dew point temperature can also be the air saturation temperature at the same pressure and humidity ratio as the considered mixture.

Due to the moisture present in the air, the heating and cooling processes undergone by the air are accompanied by changes in its relative humidity and/or humidity ratio. Consequently, the design of air processing to meet certain requirements regarding its temperature and moisture must take into consideration the heat and mass transfer associated with these state changes. Moreover, in certain cases (such as cooling below the dew point temperature at constant humidity ratio), water vapor condensation will result and liquid water will leave the air.

10.1.1 Human Comfort

The main parameters that define the air quality in a space are temperature, humidity, air purity (involving dust, microorganisms, and odors), thermal radiation, and air motion.

The factors that influence the subjective sensation of comfort in a space are of extreme variety and depend on sex, age, the nature of the human activity in that space, etc. A 100% satisfaction regarding comfort is an unrealistic objective due to the previously mentioned considerations. This is why the American Society of Heating, Refrigerating and Air Conditioning Engineers Comfort Standard 55 defines an environment as thermally acceptable if at least 80% of the occupants consider it acceptable from the thermal standpoint.

The normal temperature of the human body is approximately 36.9°C and is maintained at this constant value by physiological mechanisms. If the temperature and humidity of the environment allow these mechanisms to accomplish their duty, then the subjective perception of comfort is obtained.

There are basically two ways by which the human body maintains its temperature at a quasi-constant level:

1. Metabolism, which increases or decreases the rate of internal heat production as the body temperature decreases or increases, and which depends on the level of effort that characterizes the human activity.

2. Control of the heat transfer between the body and the environment: If the external temperature is low, the body's heat losses are reduced by reducing the intensity of blood circulation (vascular constriction at the skin level). If the exterior temperature is too high, blood vessels that irrigate the skin dilate, allowing a higher blood flow and consequently a higher rate of heat rejection. This cooling mechanism is combined with the heat rejection by means of the latent heat of evaporation of the sweat.

The heat release of the human body can vary from 75 (sleep) to 700 W (sustained hard work), in accordance with the degree of effort of the activity. The corresponding limits of the humidity release by perspiration are 0.05 and 0.45 kg/h, respectively.

The relative humidity of the air is not directly perceived as such, but it contributes to the general sensation of comfort by encouraging or hampering sweating. This is why a higher temperature is endurable when air is dry, whereas moderate temperatures seem intolerable if the relative humidity approaches 100%.

In all cases, the air motion is of great importance since it intensifies the convection heat transfer of the body. An air velocity of approximately 0.25 m/s is acceptable from the standpoint of thermal comfort. Another factor that determines the thermal comfort is thermal radiation, which becomes important if there are surfaces in the environment whose temperatures are different from the air temperature.

These factors act in strong correlation with each other, which is why their influence must always be considered in an overall manner. For example, Figs. 17 and 18 present the combined action of air temperature and relative humidity and air temperature and velocity, respectively, in defining the comfort domain.

The air-conditioning systems are classified according to the following criteria:

Application: Comfort and technological

Air pressure: High and low (classical)

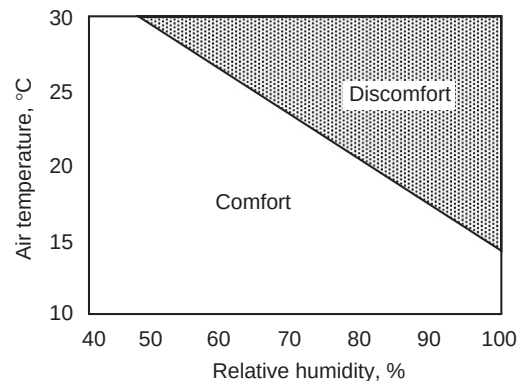


FIGURE 17 The comfort and discomfort domains resulting from the combined action of the air temperature and relative humidity.

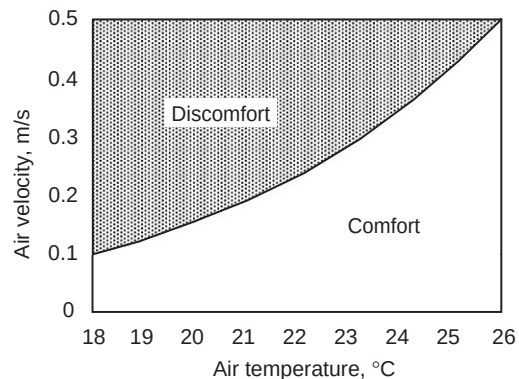


FIGURE 18 The combined action of air velocity and temperature in defining the comfort and discomfort domains.

Flow rate: Constant and variable

Heat transfer agent: All air, air and water, or all water.

The design of an air-conditioning system must consider many variables that determine the air parameters, such as solar radiation, air infiltration, heat gains due to conduction and convection from boundaries and internal objects, and latent heat and humidity gains caused by water-containing objects. All these heat and moisture sources determine the cooling load, which must be equal to the heat rate that has to be removed from the space by the air-conditioning equipment. This requirement is rarely met since the cooling load is almost permanently variable as a consequence of the ever-changing conditions (season, solar radiation, ambient temperature, wind, occupancy of the space, daily and weekly work cycle, variable operation of the heat and moisture-generating equipment, etc.). Therefore, it is recommended that the air-conditioning system should be designed for a peak cooling load and it will be operated intermittently at loads less than the peak (design) value.

The indoor design conditions are 25°C and 50% relative humidity. The necessity of a good match with the cooling load has determined the design of packed refrigeration systems of different refrigeration powers, which are delivered ready to be integrated in the air-conditioning systems:

- VCR systems equipped with reciprocating compressors: power range, 11.6–464 kW
- VCR systems with centrifugal compressors: power range, 348–2320 kW
- Lithium bromide–water absorption refrigeration systems: power range, 116–5800 kW
- Steam jet refrigeration systems: power range, 10–6000 kW.

10.2 Temperature and Humidity Requirements for Different Activities

The function of air-conditioning systems is to control the temperature and humidity of a space in order to either meet the comfort requirements of the personnel in that space or create the microclimate in an industrial space. The thermal comfort conditions for a space impose a temperature within the range 20–25°C, and the relative humidity must be in the range 35–65%. Many industrial processes involve technological prescriptions that impose special temperature and/or humidity values for the air, in some

TABLE VII

Temperature and Relative Humidity Requirements for Some Activities

Domain	Activity	Temperature (°C)	Relative humidity (%)
Libraries	Book storage	18–25	40–50
Museums	Painting galleries	18–24	40–55
Beverages	Brewing	10–15	80–85
Rubber	Vulcanization	26–28	25–30
	Surgical material	24–33	25–30
Chocolate	Packaging	18	55
Mechanical	Precision assembly	20–24	40–50
Pharmaceuticals	Manufacturing	21–27	10–50
Tobacco	Tobacco storage	32	85–88
	Cigarette manufacturing	21–24	55–65
	Packaging	23	65
Textiles	Cotton spinning	24–27	50–65
	Wool spinning	24–27	60–80

cases in very narrow ranges. Table VII illustrates these requirements for some of the most common domains.

10.3 Basic Central Air-Conditioning

This is a single-zone air-conditioning system and supplies a constant or variable air volume at low, medium, or high pressure. Usually, the system is located outside the air-conditioned area.

The typical components of air-conditioning equipment are air filters; a refrigeration unit; cooling coils; heating coils using steam, hot water, or electricity; an air humidification unit; air ducts; water pipes; fans; and control equipment.

Usually, central systems supply to an entire building both warm air for heating during the cold season and cold air for cooling in the warm season. The outdoor air undergoes a series of heating and cooling processes combined with humidification and/or drying, with a resulting stream of processed air that meets the comfort requirements. Only the air-conditioning subsystems that comprise the general system are discussed further.

The air-handling components of such systems can be arranged in various configurations according to the specific requirements of the space to be air-conditioned. The air-conditioning system must be easy to maintain and must accomplish its task in the simplest configuration possible. The central systems

may be classified as all-air systems, air-and-water systems, and all-water systems.

10.3.1 All-Air Systems

These systems supply only air to the conditioned space and process it in order to offset the maximum cooling load from the building. For spaces in the building whose cooling load is less than the maximum one, a reheat of the air can be provided by adequate heating equipment.

A typical all-air system is shown in Fig. 19. All-air systems are of two types: single-path and dual-path. In single-path systems, a single stream of air flows through the heating and cooling coils that are arranged in series. In dual-path systems, the heating and cooling coils are placed on separate paths and the processed air stream results by blending the heated and cooled streams.

All-air systems are used in buildings that have several spaces that each require a separate control of the air's temperature and humidity, such as hospitals, office buildings, hotels, and stores. All-air systems are used wherever a close temperature and humidity control is required (e.g., hospital operating rooms, computer rooms, and textile and tobacco industries).

10.3.2 Air-and-Water Systems

These air-conditioning systems were developed and implemented after World War II due to the need to reduce the space required for the air ducts in buildings. Their typical use is in high-rise structures. A central unit processes the minimum amount of

fresh air at 14–16°C, called primary air, and distributes it under a relatively high pressure to the rooms and spaces to be air-conditioned. Due to the higher pressure, the air velocity through the ducts is high, and therefore the flow section area of the ducts is small, as is the space required. As a result, higher power fans are required. Water (called secondary water) is also processed in the central unit. Primary air and secondary water are distributed to terminal units in the conditioned spaces throughout the building. The room terminal may be an induction unit or a fan-coil unit. The primary air supplied by the central air-conditioning system is mixed in the room terminal units with the return (secondary) air and distributed in the rooms. The advantage of this system is that the temperature can be adjusted in each room. However, this configuration complicates the piping because each room or space must have its own heating and cooling pipes (water is the heat transfer agent), and consequently the costs are higher.

The induction unit uses the ejection effect produced by the high-pressure primary air that flows through nozzles and sucks the interior (return) air, which is forced through heat exchangers that cool or heat it as required. The induction units are usually installed under a window or in an overhead location. They have some advantages, such as individual room temperature adjustment, less space required, easy maintenance (no moving parts), long life span, and smaller size of the central unit due to the small volume of primary air to be processed. On the other

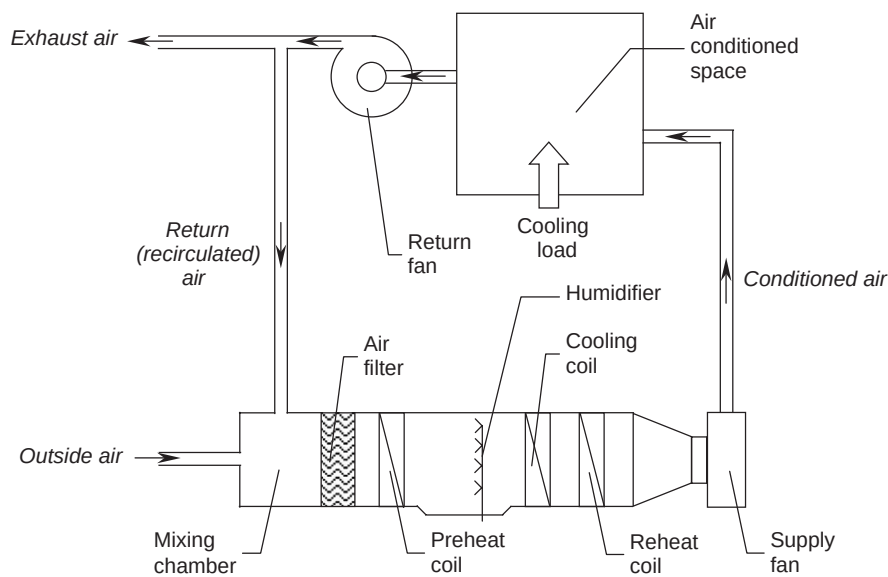


FIGURE 19 Scheme of an all-air central air-conditioning system.

hand, fouling of the secondary air side of the cooling or heating coils is possible and high energy consumption is necessary.

The typical components of a fan-coil unit are a filter, finned-tube coils, and a fan. Hot or cold water circulates through the coils. The main drawback is the noise produced by the fan. Some configurations contain an additional electric heater, which is used during the season transition to avoid changeover problems.

The air-and-water systems are used when no accurate control of the humidity is required (schools, office buildings, hotels, hospitals, etc.).

In all-water systems, the only cooling agent is water, which flows through finned coils, and individual fans blow air across the coils. These systems are recommended for their flexibility, but only air cooling and dehumidification are possible, unless a central humidifier is provided. The typical room terminal unit in this case is the fan-coil unit.

10.4 Unitary Air Conditioners

Unitary systems (also called multiple-packaged unit systems) are intended to match the air quality criteria for separate (individual) spaces. Their design allows a large variety of configurations and placements. The arrangement may be single or split, placement may be wall mounted or suspended from ceilings, cooling may be done with an evaporative condenser, they may be water or air cooled, etc. The main advantage of the unitary systems is flexibility. They operate more economically than central systems in large spaces, in which the placement of the units can cover the local cooling loads in a very effective manner.

Unitary systems (cooling capacity, 2.5–400 kW) are suitable when low initial cost, lower performance, and simple installation are required. Typical applications are office buildings, motels, and stores.

They are manufactured in different configurations in order to meet the multitude of possible requirements and applications. Each system consists of filters, heating and cooling coils, fans, one or more refrigeration compressors, and control equipment. Sometimes, multipackaged units for perimeter rooms are used in combination with central air-conditioning to increase flexibility.

Outdoor unitary systems are intended to cool or heat an entire building and consist of unitary systems, air ducts, and temperature and humidity control equipment. Usually, they are mounted on the roof, but in certain situations they can be mounted on the ground level.

Indoor unitary air-conditioning systems are located in the service areas of the building, close to the conditioned space, and are composed of an indoor unit with a water- or an air-cooled condenser, air ducts, and temperature and humidity control equipment.

The split configuration consists of two units connected by the refrigerant pipes:

1. The exterior unit (usually mounted in a console on the external face of the wall, on the ground, or on the roof), which contains the compressor, a fan, and the condenser
2. The interior unit, which consists of the evaporator, the reheat coil, the humidifier, and the fan

The advantages of unitary air conditioners are as follows:

- Simple temperature control by the room occupant
- One room can be heated or cooled independently from other rooms in the building
- Malfunction affects only one room
- Low initial cost
- Space economy
- Individual energy metering possible
- Immediately available.

Disadvantages include:

- Limited performance options and air distribution control
- No possibility of humidity control
- Higher energy consumption due to the usually lower efficiency

10.5 Room Air Conditioners

Room air conditioners (cooling capacity, 1.2–10.5 kW) are intended for air cooling, humidifying, and cleaning and are designed for mounting in windows, as consoles, or through special openings in walls. They have low installation costs and a relatively low price, are reliable and easy to maintain, and no or a minimum number of air ducts are required.

The window-mounted air-cooling unit (Fig. 20) only cools the air, without controlling its relative humidity. The refrigeration power of such units varies from 1 to 8 kW. The condenser is air cooled and the fan blows the air through the evaporator and the condenser. Window-mounted room air conditioners are usually used when low cost, simplicity of

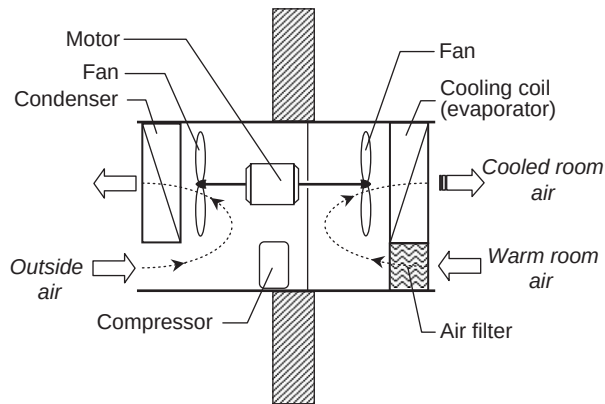


FIGURE 20 Typical window-mounted room air conditioner.

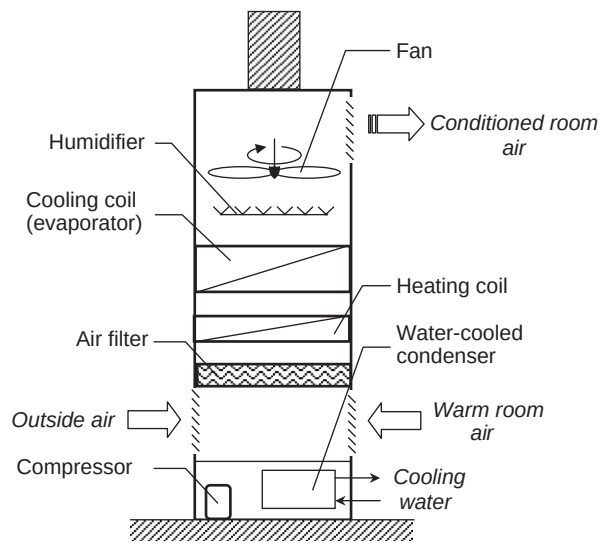


FIGURE 21 Typical through-the-wall room air-conditioning unit.

operation, and maintenance are primary criteria in selecting a candidate system (motels, low-cost apartments, and hotels), as auxiliaries to central systems, or as a backup for certain spaces when the basic system is down.

Through-the-wall mounted air conditioners (Fig. 21) contain a complete air-cooling and air-handling system packaged in one unit. They are used in buildings that require several temperature control zones (office buildings, hotels, and schools) and because less space is required compared with a central system, installation is relatively simple, and initial costs are generally lower than those of a central system. However, their noise level generally limits their applications.

SEE ALSO THE FOLLOWING ARTICLES

Air Pollution, Health Effects of • District Heating and Cooling • Electricity Use, History of • Heat Transfer • Indoor Air Quality in Developing Nations • Indoor Air Quality in Industrial Nations • Solar Cooling, Dehumidification, and Air-Conditioning • Thermal Comfort • Thermodynamic Sciences, History of

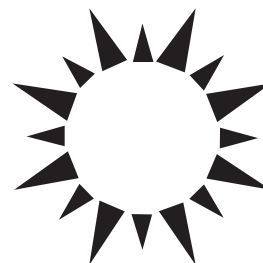
Further Reading

- Anderson, E. E. (1993). "Thermodynamics." PWS, Boston.
- Bejan, A. (1988). "Advanced Engineering Thermodynamics." Wiley, New York.
- Howell, J. R., and Buckius, R. O. (1987). "Fundamentals of Engineering Thermodynamics." McGraw-Hill, New York.
- Kutz, M. (Ed.) (1998). "Mechanical Engineers' Handbook". Wiley, New York.
- Moran, M. J., and Shapiro, H. N. (1999). "Fundamentals of Engineering Thermodynamics." Wiley, New York.
- Parsons, R. A. (ed.). (1987a). "ASHRAE Handbook." HVAC Systems and Applications. American Society of Heating, Refrigerating and Air-Conditioning Engineers, Atlanta.
- Parsons, R. A. (ed.). (1987b). "ASHRAE Handbook." Equipment American Society of Heating, Refrigerating and Air-Conditioning Engineers, Atlanta.
- Parsons, R. A. (ed.). (1990). "ASHRAE Handbook." Refrigeration Systems and Applications. American Society of Heating, Refrigerating and Air-Conditioning Engineers, Atlanta.
- Sonntag, R. E., and Van Wylen, G. (1982). "Introduction to Thermodynamics Classical and Statistical." Wiley, New York.

Remote Sensing for Energy Resources

FLOYD F. SABINS

Remote Sensing Enterprises, Inc.
Fullerton, California, United States



-
1. Remote Sensing Technology
 2. Oil Exploration
 3. Environmental Applications

Glossary

anticline An arch in sedimentary strata. Oil may occur in the crest of anticlines.

electromagnetic spectrum The continuous sequence of electromagnetic energy arranged according to wavelength or frequency.

Landsat The unmanned Earth-orbiting satellites operated by NASA that acquire repeated remote sensing images.

radar Acronym for radio detection and ranging. Radar is an active form of remote sensing that records wavelengths from millimeters to centimeters.

Remote sensing is defined as the science of acquiring, processing, and interpreting images from satellites and aircraft that record the interaction between matter and electromagnetic energy. Remote sensing images of the earth are acquired in three wavelength intervals, or regions, of the electromagnetic spectrum. The visible region ranges from 0.4 to 0.7 μm and is divided into the blue, green, and red bands. The infrared (IR) region ranges from 0.7 to 30 μm and is divided into the reflected IR and thermal IR portions. The reflected IR portion ranges from 0.7 to 3.0 μm ; the energy is predominantly reflected sunlight at wavelengths longer than visible light. The Landsat and other satellite systems acquire valuable images in the visible and reflected IR regions. The thermal IR portion ranges from 3.0 to 15.0 μm ; the energy is radiant, or heat, energy. Thermal IR images have considerable potential for oil exploration in arid and semiarid terrains; however, the method has

been underutilized in recent years largely because suitable images are lacking.

Visible, reflected IR, and thermal IR images are acquired in the passive mode by remote sensing systems that simply record the available energy that is reflected or radiated from the earth. The microwave region ranges from 0.1 to 30 cm; images are primarily acquired in the active mode called radar. An aircraft or satellite radar antenna transmits short bursts of microwave energy toward Earth. The energy interacts with materials and is returned to the antenna system, which records the information for later display as an image. Remote sensing is a rapidly evolving science with new imaging systems deployed on an irregular schedule by various nations and private concerns. The World Wide Web is a primary source of information about this dynamic science. Table I lists representative Web sites, many of which are linked to additional sites.

1. REMOTE SENSING TECHNOLOGY

The major application of remote sensing to energy is for oil exploration and, to a lesser extent, assessing the environmental impact of oil-related operations. The two most widely and successfully used systems are Landsat and radar.

1.1 Landsat Thematic Mapper Images

Landsat is an unmanned NASA satellite that orbits the earth in a sun-synchronous pattern at an altitude of 705 km. The satellite orbits over almost the entire earth in 16 days. If an area is cloudy during one overflight, suitable images may be acquired during subsequent 16-day cycles. Beginning in 1982, the

second-generation Landsats 4 and 5 carried the thematic mapper (TM), which is a multispectral scanner that records images in seven separate wavelength intervals, or bands, for each scene. Bands 1–3 record blue, green, and red light. Bands 4, 5, and 7 record reflected IR energy. Band 6 records thermal IR energy, but it is rarely used for exploration. Table II lists the bands and their characteristics. Figure 1 shows reflectance spectra of vegetation and common sedimentary rocks together with wavelengths recorded by TM bands. The spectra of the different rocks are very similar in the visible bands but have major differences in the reflected IR bands; therefore, bands 4, 5, and 7 are especially useful for mapping different rock types.

Spatial resolution of the visible and reflected bands is 30 m. The current Landsat 7 carries the enhanced thematic mapper with an additional band 8 (0.52–0.90 μm), which has a spatial resolution of 15 m and can be convolved with the other bands during computer processing to improve their apparent spatial resolution.

Figure 2 is a subscene from a TM band 4 image of central Wyoming. The subscene displays less than

0.5% of the area covered by the full scene (170 $\times$ 185 km). Figure 3 shows the major features in the image and is generalized from a detailed geologic map that was interpreted for the full scene. The image includes four anticlines, two of which (Gebo and Little Sand Draw) are oil fields. These fields were discovered prior to the launch of Landsat, but this example demonstrates the value of the images for oil exploration.

Any three TM bands may be merged in any combination of blue, green, and red to produce color composite images that are far superior to panchromatic images, such as Fig. 2. TM images of the world are available at modest prices from the U.S. Geological Survey EROS Data Center in Sioux Falls, South Dakota (Table I). The images are distributed in digital format and must be computer processed to produce images. A variety of commercial desktop computers and software are available for this purpose.

Multispectral scanners, such as TM, record up to approximately 10 bands of data with wavelength intervals on the order of 0.10 μm . Hyperspectral scanners are a special form of multispectral scanners

TABLE I
Representative Remote Sensing Web Sites^a

Web site	Web address	Comments
American Society for Photogrammetry & Remote Sensing	www.asprs.org/resources.html	Enter keyword "Satellites" for charts and tables describing past, current, and future remote sensing satellites
Canada Center for Remote Sensing	www.ccrs.nrcan.gc.ca	Canadian counterpart of EROS Data Center
EROS Data Center	edc.usgs.gov	Search and other aerial photographs, Landsat, radar, and other images
EROS Data Center DAAC	edcdaac.usgs.gov	Search and order SIR-C, ASTER, AVHRR, TIMS, MODIS, TERRA
Jet Propulsion Laboratory homepage	www.jpl.nasa.gov	Home page for JPL images: A VIRIS, TIMS, ASTER, SIR, others
JPL MASTER site	masterweb.jpl.nasa.gov	View and other MASTER aircraft hyperspectral images
JPL Shuttle Imaging Radar (SIR) site	southport.jpl.nasa.gov	SIR and Shuttle Radar Topography Mission (SRTM 2000)
NASA home page	www.nasa.gov	Links to NASA headquarters and centers
Purdue University	dynamo.ecn.purdue.edu/~biehl/multispec	Source for Multispec freeware for processing image data
Radarsat International	www.rsi.ca	Search and order information for worldwide satellite radar images
Space Imaging	www.spaceimaging.com	Search and other information for IKONOS high-resolution images
SPOT Image	www.spot.com	Search and order information for SPOT satellite images
USGS Spectroscopy Laboratory	speclab.cr.usgs.gov	Basic reference for hyperspectral imaging

^aFrom Sabins (1997), Table 1–8.

TABLE II

Landsat Thematic Mapper (TM) Spectral Bands^a

TM band, Landsat 4 and 5	ETM band, Landsat 7	Wavelength (μm)	Characteristics
1 (spatial resolution 30 m)	1 (spatial resolutions 30 m)	0.45–0.52	Blue green; maximum penetration of clear water; useful for distinguishing soil from vegetation and deciduous plants from coniferous plants
2 (30 m)	2 (30 m)	0.52–0.60	Green; matches green reflectance peak of vegetation, which is useful for assessing plant vigor
3 (30 m)	3 (30 m)	0.63–0.69	Red; matches a chlorophyll absorption band that is important for discriminating vegetation types
4 (30 m)	4 (30 m)	0.76–0.90	Reflected IR; useful for determining biomass content and for mapping shorelines
5 (30 m)	5 (30 m)	1.55–1.75	Reflected IR; indicates moisture content of soil and vegetation; penetrates thin clouds
6 (120 m)	6 (60 m)	10.40–12.50	Thermal IR; nighttime images useful for thermal mapping and for estimating soil moisture
7 (30 m)	7 (30 m)	2.08–2.35	Reflected IR; especially useful for distinguishing rock types and for mineral exploration
	8 (15 m)	0.52–0.90	Visible and reflected IR; may be convolved with other bands to improve apparent spatial resolution

^aFrom Sabins (1997), Table 3–4.

that record many tens of bands with wavelength intervals on the order of $0.01\ \mu\text{m}$. The Web sites for the Jet Propulsion Laboratory (JPL), U.S. Geological Survey Spectroscopy Laboratory, and EROS Data Center (Table I) describe the major hyperspectral systems. Hyperspectral images have had limited applications to oil exploration.

In addition to Landsat, a number of other Earth-orbiting remote sensing satellites have been launched and others are planned. The Web site of the American Society for Photogrammetry and Remote Sensing (Table I) summarizes the characteristics of these satellites.

1.2 Radar Images

Radar is an active form of remote sensing that is shown diagrammatically in Fig. 4. Pulses of microwave energy, at wavelengths of a few centimeters to a few tens of centimeters, are directed to the terrain from a fixed antenna carried on an aircraft or satellite. Energy is directed downward toward the side of the ground track. Each radar pulse “illuminates” a narrow strip, or scan line, of terrain normal to the track. The depression angle is measured between the horizon and the center of the pulse of radar energy. Figure 4 shows a single scan line. The incident energy interacts with the terrain and portions are returned to the radar system, which

records the data. Three characteristics of terrain determine the intensity of the radar return:

1. Dielectric constant, which is largely a function of moisture content.
2. Surface roughness at a scale of centimeters. Smooth surfaces, such as calm water, have dark signatures. Rough surfaces, such as forests and coarse gravel, have bright signatures.
3. Topography at a scale of meters to kilometers. Figure 4 shows that slopes facing toward the radar antenna have bright signatures called highlights. Slopes facing away from the antenna have dark signatures called shadows.

Radar images have two major advantages over visible and reflected IR images, such as Landsat. Because radar is an active system that provides its own source of energy, images can be acquired day or night and through clouds, rain, and fog. Aircraft radar images are typically acquired with a low depression angle of 20° or less (Fig. 4), which causes strong highlights and shadows from topographic features that may express geologic features. Spatial resolution of modern synthetic aperture radar (SAR) images ranges from meters to tens of meters, even at satellite altitudes. Topographic maps are produced from repeated images using the process called interferometric SAR (IFSAR), which is described on the JPL Shuttle Imaging Radar Web site (Table I).

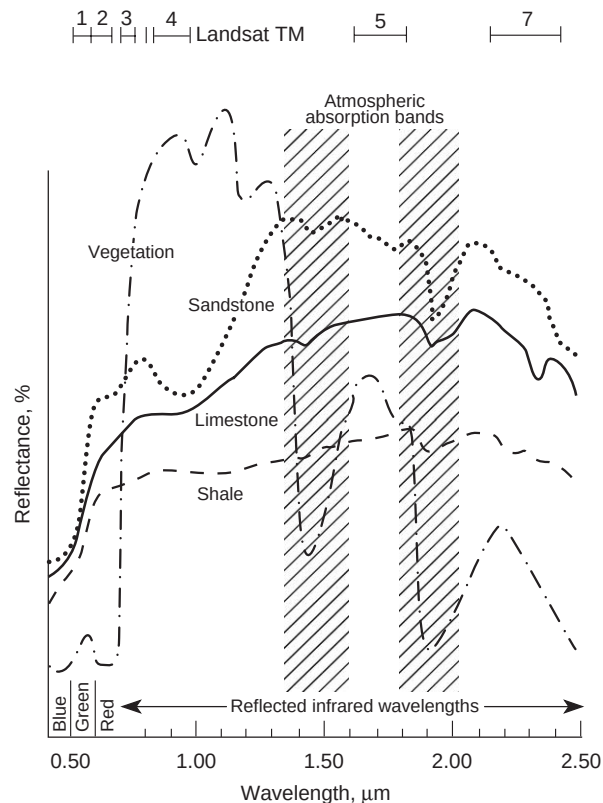


FIGURE 1 Reflectance spectra of vegetation and sedimentary rocks. Positions of TM bands are shown. Note that rock spectra have more distinguishing characteristics in the reflected IR region than in the visible region. Atmospheric absorption bands are caused by water vapor in the atmosphere, which prevents image acquisition. From Sabins (1997, Fig. 3-1).

Figure 5 is a radar image along the south flank of the Papuan Fold and Thrust Belt in Papua New Guinea (PNG). Figure 6 is a generalized geologic map interpreted from the image. The terrain consists of northwest-trending anticlines and faults that form topographic ridges and scarps. Radar illumination was toward the north, which forms bright highlights on southwest-facing slopes and dark shadows on northeast-facing slopes.

The Web sites for the EROS Data Center and JPL provide examples and information on aircraft and satellite radar images acquired by U.S. government agencies. The Radarsat Web site provides access to satellite radar images acquired by Canada.

1.3 Digital Image Processing

Except for conventional aerial photographs, all remote sensing images are recorded in digital format and are computer processed to produce images for



FIGURE 2 Landsat TM band 4 (0.76–0.90 μm) image of subsurface in central Wyoming. This subsurface (190 km^2) is taken from a TM scene that covers 31,450 km^2 . From Sabins (1997, Fig. 3-7D).

interpretation. The three functional categories of image processing are

1. Image restoration, which compensates for data errors, noise, and geometric distortions introduced during the scanning, recording, and playback operations.
2. Image enhancement, which alters the visual impact that the image has on the interpreter in a manner that improves the information content.
3. Information extraction, which utilizes the decision-making capability of the computer to recognize and classify pixels on the basis of their digital signatures.

The Purdue University Web site (Table I) provides image processing software.

2. OIL EXPLORATION

The major energy application of remote sensing is for oil exploration. In the early stages of an exploration

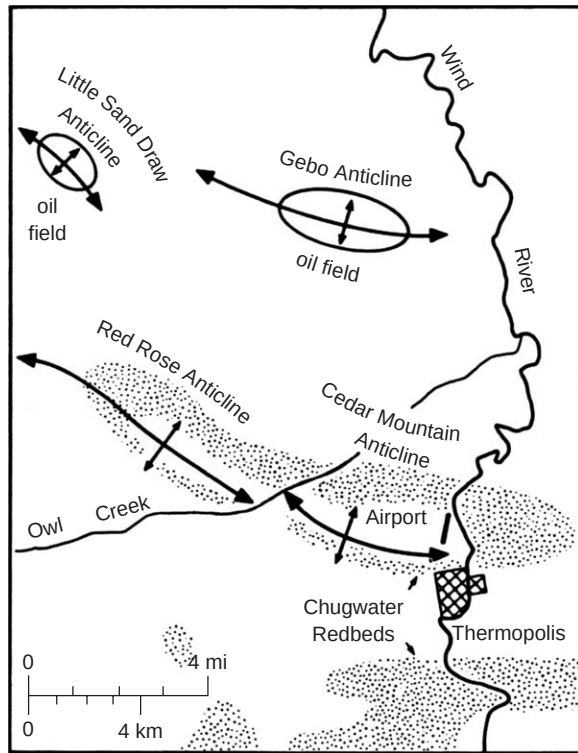


FIGURE 3 Map showing major features in Wyoming subsense. From Sabins (1997, Fig. 3-7H).

project, images are used to outline sedimentary basins and potential drilling prospects. Geologic fieldwork and geophysical surveys are used to verify and refine the prospects prior to drilling exploratory wells. In 1978, the Chevron Corporation (now ChevronTexaco) utilized images from an early version of Landsat to help discover the first oil fields in the Sudan. Remote sensing has subsequently played a key role in major discoveries in PNG and the Central Arabian Arch. The PNG project area is a tropical rain forest with complex geology, whereas the Arabian project is a desert with relatively simple geology. These two case histories illustrate the successful application of different remote sensing technologies in widely different geologic and environmental settings.

2.1 Papua New Guinea Project

In 1984, the Chevron Corporation explored two petroleum prospecting licenses (PPLs) in the eastern portion of the island of New Guinea, which is part of the nation of PNG. The PPLs cover an area of 14,400 km² that spans the boundary between the Fly-Strickland Lowlands in the southwest and the Papuan Fold and Thrust belt in the northeast. Oil

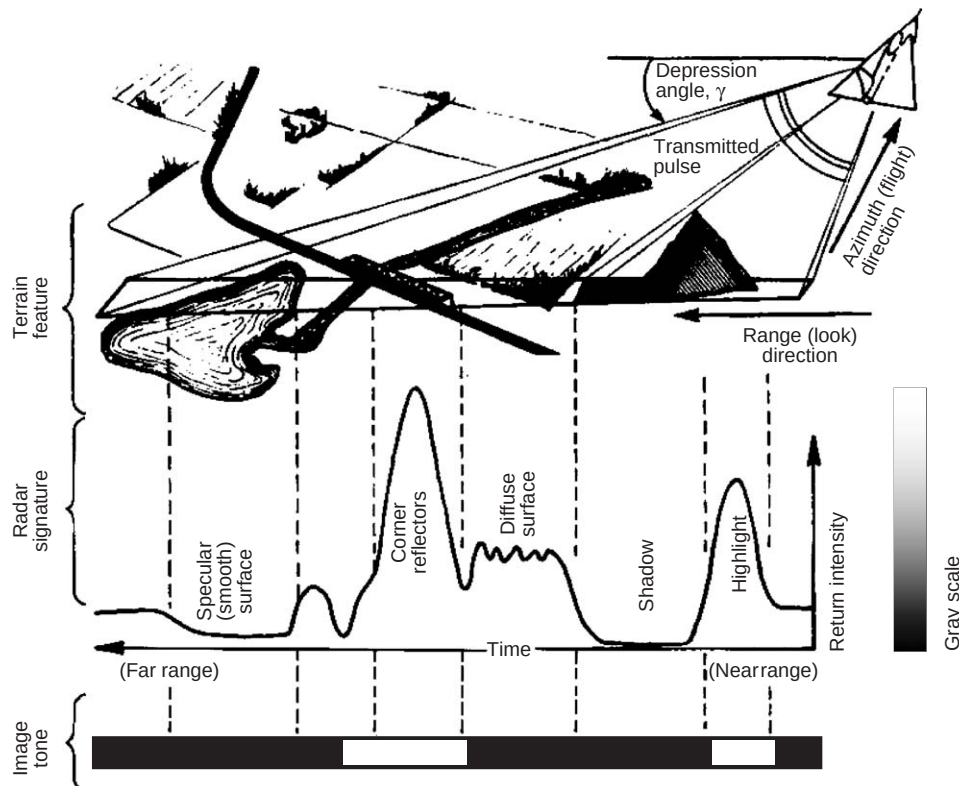


FIGURE 4 Terrain returns and image signatures for a pulse of radar energy. From Sabins (1997, Fig. 6-3).

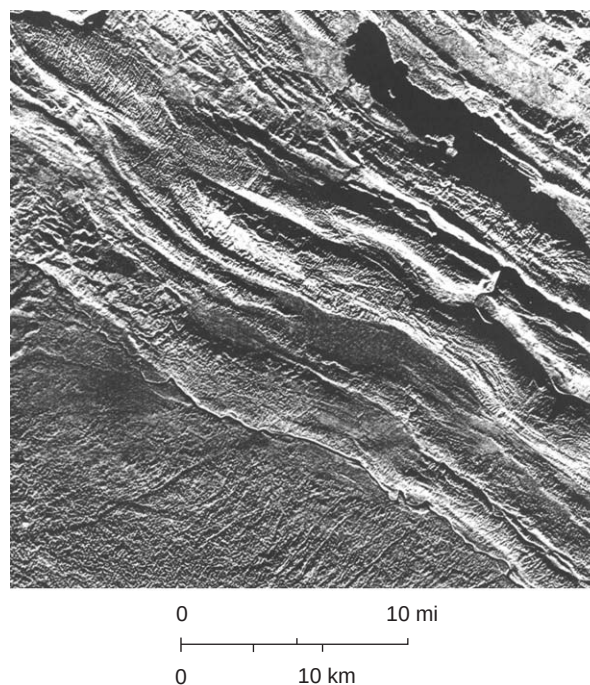


FIGURE 5 Aircraft radar image of the Lake Kutubu area of Papua New Guinea. From Sabins (1997, Fig. 10-8).

companies have explored and drilled the region for many years because of the numerous surface seeps of oil and gas and hydrocarbon shows in exploratory wells. When this project began, 144 wells had been drilled in PNG, but no commercial oil fields had been discovered. This remote area has a canopy of tropical rain forest and a hot, humid climate with persistent cloud cover and heavy rainfall. The terrain is dominated by steep ridges and deep river channels. Much of the region is underlain by limestone that weathers into steep pinnacles and pits, called karst topography, that cannot be traversed by vehicles. Reliable seismic data were unobtainable because of the rugged terrain and the weathering effects caused by pervasive subsurface solution cavities.

At the beginning of the project, Chevron decided to rely on radar images and limited fieldwork. A contract radar survey aircraft covered the area with a series of parallel image strips that were compiled into mosaics and interpreted to produce geologic maps. Based on these maps, the first well was drilled on a promising structure. The well was dry but had encouraging shows of oil. Exploration then shifted southeast to the region of Lake Kutubu. Figure 5 is

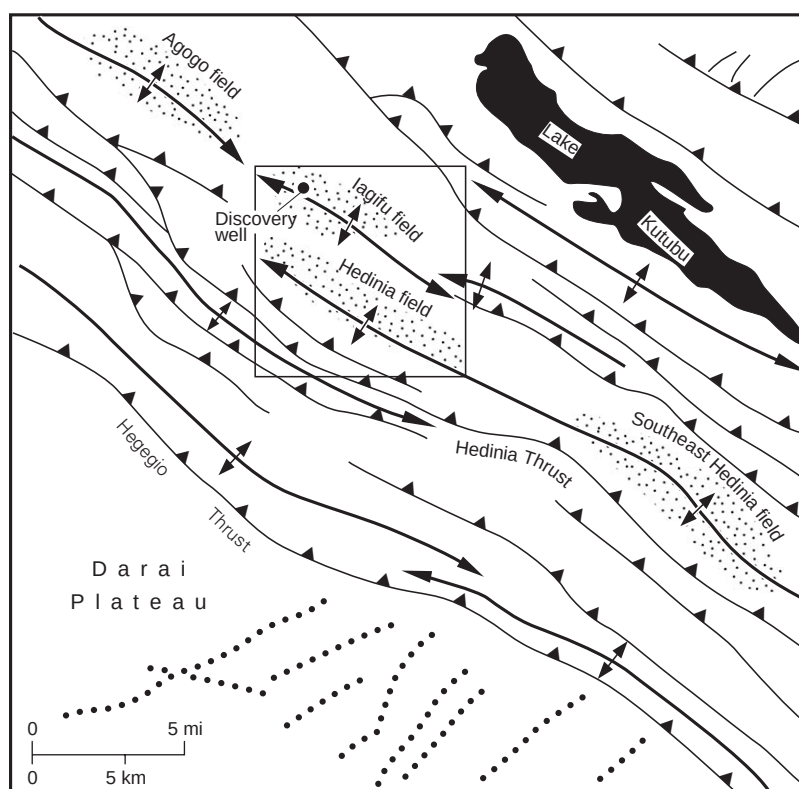


FIGURE 6 Geologic interpretation of radar image of the Lake Kutubu area of Papua New Guinea. From Sabins (1997, Fig. 10-9).

the radar image and Fig. 6 is the interpretation map for the area. Based on the radar interpretation and limited fieldwork, the first commercial oil discovery in PNG was made in March 1986 on the Iagifu Anticline at the location shown in Fig. 6. Subsequently, Chevron has discovered several additional commercial fields. A number of promising structures remain to be drilled in the region.

2.2 Central Arabian Arch Project

For a number of years, the Arabian American Oil Company (Aramco) was restricted to explore only within the “Retained Areas” of existing production that are located along the Arabian Gulf. Once the giant fields were drilled, no new fields were discovered through 1986. At that time, the government directed the company, now called Saudi Aramco and owned by the government, to explore the vast region of the Kingdom with oil potential that lies outside the Retained Areas. In early 1988, Saudi Aramco asked the Chevron Remote Sensing Research Group to conduct a remote sensing analysis of this region. The initial phase of the study covered the Central Arabian Arch, an area of approximately 200,000 km². The two objectives of the project were to map the regional geology and identify surface features that may be the expression of oil prospects.

The climate, topography, and geology of Saudi Arabia are the opposite of those in PNG. The arid climate and lack of vegetation are ideal for using Landsat images. The geologic structure is simple and uniform, in contrast to the complex faulting and folding in PNG.

Landsat TM bands 2, 4, and 7 were combined in blue, green, and red to produce color images that were interpreted to produce regional geologic maps at a scale of 1:250,000 (1 in. = 4 miles). Aside from the sand seas, bedrock is well exposed. Geologic structure consists of regional dips of approximately 1° to the northeast or east on opposite flanks of the Central Arabian Arch. Distinctive color signatures on the images enabled the interpreters to divide several formations of sedimentary rocks into two or more mappable subdivisions, which proved valuable for identifying oil prospects.

Identifying potential drilling prospects was challenging because the geologic structure is dominated by broad uniform dipslopes that extend for several hundred kilometers along the strike and several tens of kilometers in the downdip direction. Unlike PNG, there are no obvious folds. Within the Retained Areas, oil fields are trapped in gentle anticlines

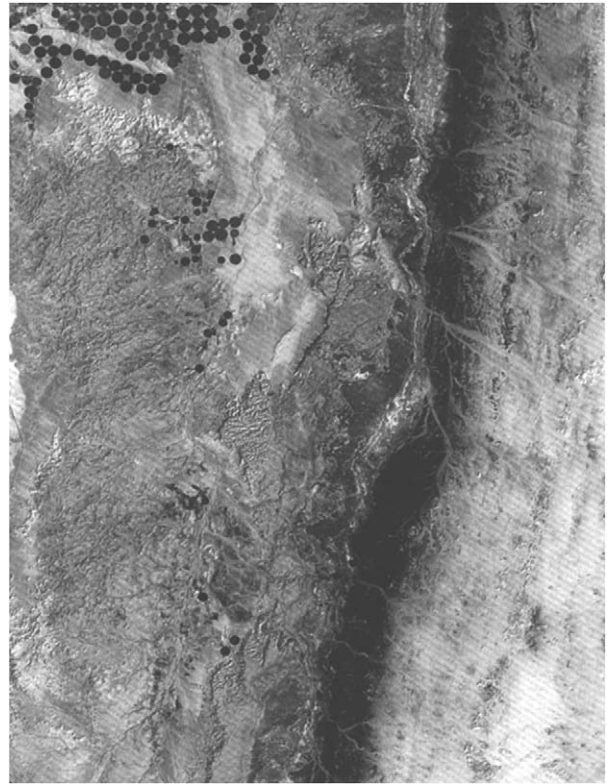


FIGURE 7 Landsat TM image of the Raghib prospect, Central Arabian Arch. From Sabins (1997, Plate 20-D).

formed above high-angle faults that offset the underlying rocks. The anticlines grade upward into flattening of the regional dip to form subtle structural terraces at the surface. We used this information to develop a three-dimensional geologic model for recognizing potential oil prospects on the images. A number of prospects were interpreted and were checked by geologists in the field. Saudi Aramco then followed up with seismic surveys of the most promising prospects.

Figure 7 is the TM image of the Raghib prospect located 110 km southeast of Riyadh. This black-and-white version lacks much of the information shown in the original color image that was used for our interpretation. Figure 8 is the geologic interpretation map. Jurassic and Cretaceous strata dip 1° east. Arcuate scarps form the downdip margin of an anomalous topographic depression that is mantled by windblown sand. Figure 9 is a subsurface structure map from a seismic survey that defines a northwest-trending anticline that is 35 km long and 10 km wide. The axis of the anticline is annotated on the Landsat interpretation map (Fig. 8). The topographic depression on the image coincides with the

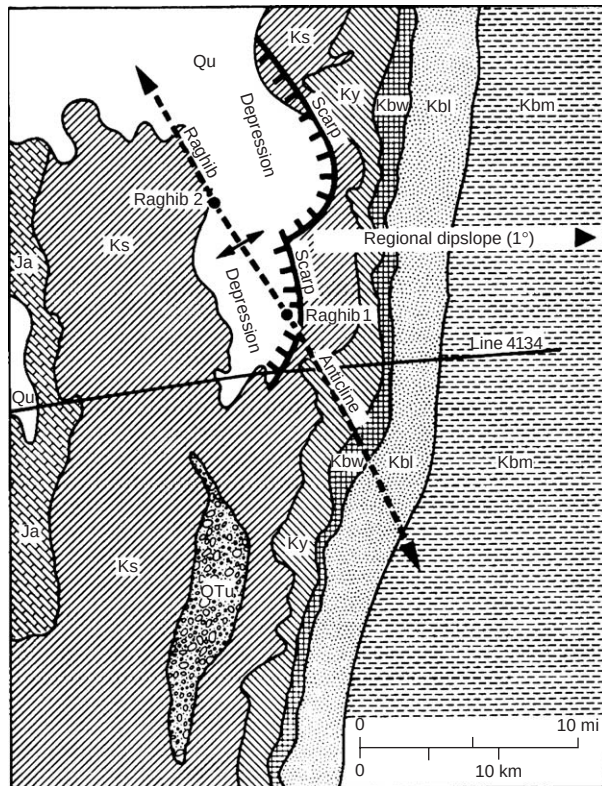


FIGURE 8 Geologic interpretation of the Raghib prospect. Raghib 1 is the discovery well and Raghib 2 is the confirmation well for the Raghib oil field. The patterns and symbols designate geologic formations. From Sabins (1997, Fig. 10-20).

crest of the subsurface anticline. In late 1989, Saudi Aramco drilled the Raghib 1 exploratory well located on the crest of the subsurface anticline (Fig. 8) and discovered a new oil field. A number of prospects remain to be explored.

2.3 Operational Applications

In addition to defining prospects, images are a valuable asset to field operations in regions that lack reliable maps. The cost and effectiveness of seismic surveys can be improved by information on topography and trafficability of the survey area. In the Sudan, Landsat images provided the only reliable base maps. The images were also digitally processed to map the different terrains, such as rivers, marshes, uplands, and papyrus swamps, that required different vehicles for the seismic surveys. In the rugged terrain of Yemen, topographic maps were compiled from satellite images and used to lay out the optimum pattern of seismic survey lines. In older oil fields, well sites and pipelines may be

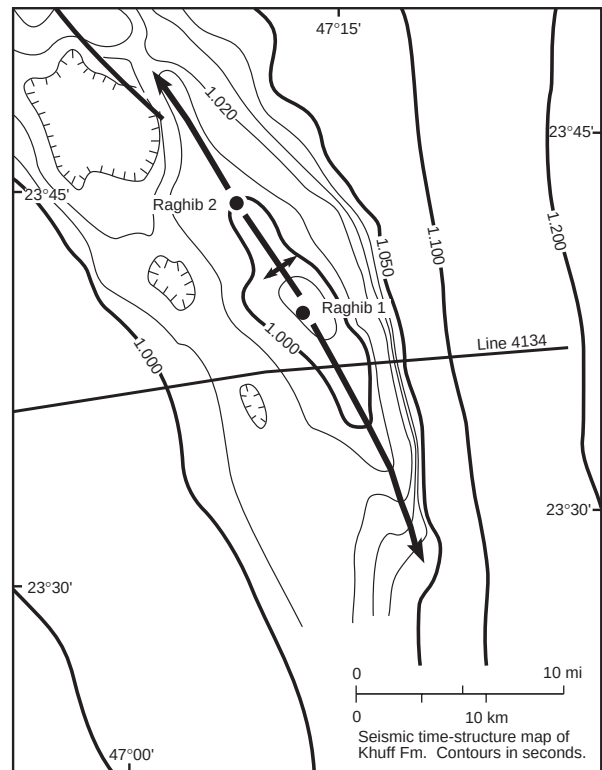


FIGURE 9 Seismic structure map of the Raghib oil field. The area coincides with the image (Fig. 7) and geologic map (Fig. 8). Contours are drawn on top of the Khuff Limestone (Permian age) and delineate a closed anticline that coincides with the surface feature. From Sabins (1997, Fig. 10-21).

mapped inaccurately but can be correctly located with high-resolution satellite images. The cost is much less than that of a conventional survey on the ground.

2.4 Direct Detection of Oil

In PNG, Saudi Arabia, and elsewhere, remote sensing is used to recognize the surface expression of the subsurface structures (folds, faults, and domes) in which oil is trapped. Oil is not detected. In a number of oil deposits, some oil leaks to the surface to form oil seeps. During the early days of the oil industry, exploratory wells were commonly located at seeps. In a few fields the leaking oil has interacted chemically with the surface rocks to alter their mineralogy and color. Some investigators have attempted to identify seeps and altered rocks as an exploration method called direct detection. To date, the results of this method have been inconclusive.

3. ENVIRONMENTAL APPLICATIONS

Offshore oil spills are predominantly caused by tanker accidents and, to a lesser extent, drilling and production activities. Remote sensing images are used to detect spills and to monitor and direct cleanup operations. Table III lists the interactions between oil and electromagnetic energy at wavelength regions ranging from ultraviolet (UV) to radar. Figure 10 shows these interactions diagrammatically. The following sections describe the detection of oil spills on remote sensing images acquired in the various spectral regions. In offshore areas, oil may leak from deposits beneath the seafloor and form slicks on the water. Remote sensing of these natural slicks has been used to aid offshore exploration.

3.1 Ultraviolet Images

The UV spectral region spans wavelengths from 0.03 to 0.40 μm . Wavelengths shorter than 0.30 μm are absorbed by the atmosphere; therefore, only wavelengths from 0.30 to 0.40 μm are used in remote sensing of the earth. Figure 10A shows that incoming UV radiation from the sun stimulates oil to fluorescence at wavelengths from 0.30 to 0.45 μm (long-wavelength UV to visible blue). UV images are the most sensitive remote sensing method for monitoring oil spills and can detect films as thin as 0.15 μm . Daylight and very clear atmosphere are essential to acquire UV images. UV energy is strongly scattered by the atmosphere, but useable images can be acquired from altitudes below 1000 m. Cameras and scanners are passive systems that operate in daytime to record fluorescence stimulated by the sun. Active UV systems have been developed for aircraft use. A laser irradiates the water with UV energy that stimulates any oil to fluoresce. The fluorescence is recorded as a spectrum or as an image. The spectrum

can be compared with a reference library of spectra to identify the oil in the spill. These active systems can acquire images both day and night but must operate at low altitudes under clear conditions.

3.2 Visible and Reflected Infrared Images

The visible spectral region spans wavelengths from 0.40 to 0.70 μm and is detectable by the eye. The reflected IR region extends from 0.40 to 3.00 μm , which is beyond the sensitivity range of the eye. The interaction between oil and electromagnetic energy in the visible and reflected IR regions is determined by the absorption and reflection of sunlight (Fig. 10B). Various types of aerial photographs record visible wavelengths and the shortest IR wavelengths (0.70–0.90 μm). Landsat TM records three visible and three reflected IR bands in digital format that is suitable for computer processing. During the invasion of Kuwait in January 1991, Iraqi forces deliberately released 4–6 million barrels of crude oil into the Arabian Gulf, where it covered approximately 1200 km^2 of water and 500 km of the Saudi Arabian coastline. This spill is the largest on record; for comparison, the 1989 Exxon Valdez spill in Alaska (<300,000 barrels) was an order of magnitude smaller. Repeated TM images tracked the extent and movement of the Kuwait spill.

3.3 Thermal IR Images

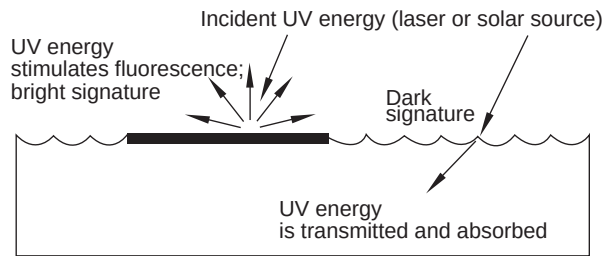
The thermal IR region spans wavelengths from 3.0 to 14.0 μm (Table III) and consists of energy radiated, or emitted, from materials. Figure 10C shows the interaction of thermal IR energy with oil and water. Both liquids have the same kinetic temperature of 18°C because they are in direct contact. Remote sensing systems, however, record radiant temperature. Radiant temperature differs from kinetic temperature because of the property called emissiv-

TABLE III

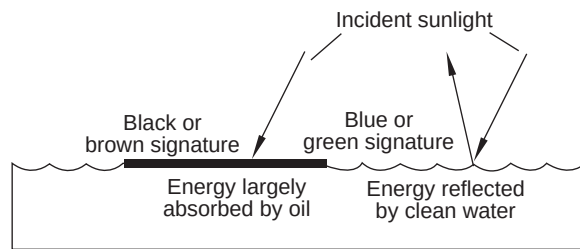
Remote Sensing Interactions and Systems for Detecting Oil on Water^a

Remote sensing region	Wavelength	Oil property detected	Oil signature	Water signature
Ultraviolet	0.3–0.4 μm	Fluorescence stimulated by sunlight or laser	Bright	Dark
Visible and reflected infrared	0.4–3.0 μm	Solar reflectance	Brown, black	Blue, green
Thermal infrared	3.0–14.0 μm	Radiant temperature, controlled by emissivity	Dark (cool)	Bright (warm)
Radar	3.0–25 cm	Dampening of capillary waves	Dark (smooth)	Bright (rough)

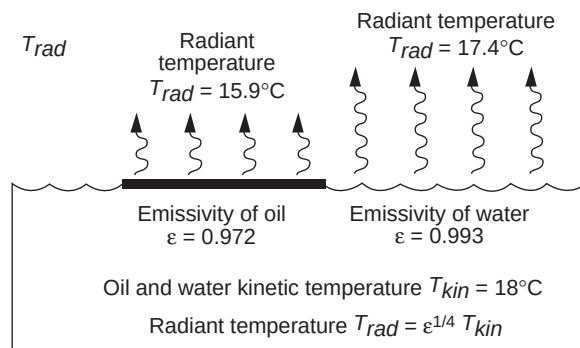
^aFrom Sabins (1997), Table 3–4.



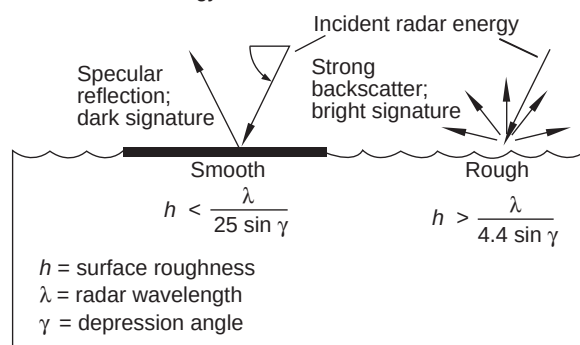
A. UV energy



B. Visible and reflected IR energy



C. Thermal IR energy



D. Radar energy

FIGURE 10 Interaction mechanism between water and oil spills. From Sabins (1997, Fig. 9-30).

ity, which expresses the ability of a material to emit radiant energy. Figure 10C shows emissivity values for oil (0.972) and water (0.993) and the equation for calculating radiant temperature. Substituting these values into the equation (for a kinetic temperature of 18°C) results in a radiant temperature of 17.4°C for water and 15.9°C for oil. The difference of 1.6°C in radiant temperature between oil and water is readily measured by thermal IR systems, which are typically sensitive to temperature differences of 0.1°C . Thermal IR images acquired from aircraft and satellites record oil slicks with cool signatures that are distinct from the warmer signatures of clean water.

3.4 Radar Images

The interaction between the surface and incident radar energy is very sensitive to variations in roughness of the surface, measured in millimeters and centimeters. Rough surfaces scatter large amounts of energy back to the antenna and have bright signatures. Smooth surfaces reflect the energy away from the antenna and have dark signatures. In Fig. 10D, clean water has characteristic ripples and capillary wavelets that cause a bright radar signature. An oil spill dampens the waves and causes a smooth surface, or slick, with a dark radar signature.

SEE ALSO THE FOLLOWING ARTICLES

Crude Oil Spills, Environmental Impact of • Gas Hydrates • Oil and Natural Gas Exploration

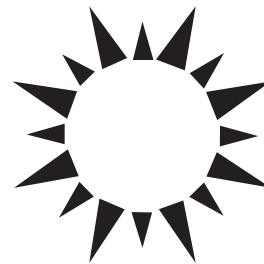
Further Reading

- Berger, Z. (1994). "Satellite Hydrocarbon Exploration." Springer-Verlag, Berlin.
- Jensen, J. R. (1986). "Introductory Digital Image Processing—A Remote Sensing Perspective." Prentice-Hall, Englewood Cliffs, NJ.
- Prost, G. L. (1993). "Remote Sensing for Geologists." Gordon & Breach, Lausanne, Switzerland.
- Sabins, F. F. (1997). "Remote Sensing—Principles and Interpretation." 3rd ed. Freeman, New York.

Renewable Energy and the City

PETER DROEGE

University of Sydney DEM Australia/World Council for Renewable Energy
Sydney, Australia



-
1. Energy Technology and Modern Urbanization
 2. Historical Perspectives
 3. Innovation in Urban Renewable Energy Technology
 4. The Urban Renewable Energy Portfolio
 5. Climate Stability in Urban Development
 6. Urban Renewable Energy Policy and Practice Examples

Glossary

city An urbanised area managed and represented by one or several local governments, culturally and communally understood as a city, with specific administrative and political boundaries.

city region The general urbanized, ecological, and economical area constituting and surrounding one or several urban nuclei, all or any of which may be defined as a city.

climate-stable practice A city's climate-stable practice is a practical commitment to lower greenhouse gas emissions by a target date such as the year 2050 to an amount that is proportionally in keeping with the globally sustainable level.

carbon-dioxide-equivalent (CO₂-e) Expresses the presence of all effective greenhouse gases as an amount that would be required in CO₂ to achieve the same effect.

greenhouse gas (GHG) Human activity effected gases that trigger global warming are carbon dioxide (CO₂), methane (CH₄), nitrous oxide (N₂O), chlorofluorocarbons, especially CFC-11, HCFC-22, and CF₄. In this article their emission levels are expressed as CO₂-e.

solar In its wider definition connotes all aspects of energy sources, which can be traced to the action of the sun. These include solar thermal, photovoltaic, biomass, bio energy, wind and wave energy, and, strictly speaking, also fossil fuels. In the pure sense, solar describes renewable forms of energy that neither create greenhouse gas emissions nor nondegradable or toxic waste.

urban(ized) area A builtup space of higher density and production and consumption functions that is distinct from rural or undisturbed natural areas.

Cities are intensive centers of social activity and the central feature in the economical technology of civilizations. Cities directly express the energy resource framework and infrastructure available at any given time in history. Global acceleration in city formation and urban growth can be ascribed to a number of technological innovations, none more profound in their impact than the rise and spread of fossil fuel resources, products, systems, and tools. The resulting massive threats—CO₂ emissions—induced climate change and the inevitably approaching plateau in worldwide fossil fuel production—raise the vexing issue of cities' overwhelming dependence on oil and other hydrocarbon energy sources as well as the question of how and when a postfossil fuel urban era can commence and whether this necessary historical transformation can take place in a stable manner.

1. ENERGY TECHNOLOGY AND MODERN URBANIZATION

The age of industrialization came into full force through the modern exploration and use of fossil fuels. As one of its most striking phenomena, the rapid expansion of cities throughout the late 19th and the 20th centuries was a direct outcome of the fossil fuel economy as well. The development and operation of cities and urbanized areas presently absorbs roughly three-quarters of the world's fossil fuel production, including intra- and interurban transport. This indicates cities' massive role in the worldwide carbon consumption and emission regime: fossil fuels, prime source of anthropogenic CO₂ emissions, supply 85% of total global commercial energy use—and their consumption is increasing at a rapid rate.

Yet economic regions, nations, and cities worldwide are under great pressure to find alternative

sources. Cities will have to be powered differently within the lifetime, prompted by a historical coincidence of carbon emissions-induced climate change, the great likelihood of an impending plateau in fossil fuel production—while energy demand continues to rise dramatically, prompting steep price rises—and mounting environmental damage due to the use of oil and coal. The use of renewable and distributed micropower systems is already on the rise, but the speed of change is seen as too low by many to meet global goals in time to avert serious crises. Combined with large-scale facility efficiency and energy conservation programs, the introduction of renewable energy generation technologies in cities offers one set of profound opportunities as well as challenges, and the reform of transport systems, settlement patterns, and land-use practices provides another.

Cities, towns, and other urban communities are regarded as settings for coordinated efforts aimed at global renewable energy technology introduction and carbon emissions reduction programs. However, substantial organizational and cultural barriers militate against immediate, wider change. Among these loom large the subsidiary regard in which cities are held in the traditional hierarchical frames of the international arrangements that deal with global issues. Another challenge is the energy-blind nature of contemporary physical planning, typically very short planning horizons and political uncertainties that prevail on the local level. In terms of policy development, measurement techniques, and planning reality, an extraordinary gulf exists between the global nature of greenhouse gas impacts and fuel depletion prospects and the local reality that represents both final impact and original source of globally experienced change.

Due to their fundamental fossil fuel reliance, most capital outlays in new buildings, infrastructure, and other urban facilities risk early obsolescence: urban areas, infrastructure systems, and individual buildings and other facilities that are based on fossil fuel construction, maintenance, and operation are likely to be rendered dysfunctional well before the end of their life. All basic urban communication infrastructures, both traditional (such as roads, rail, air, and seaports) and advanced (such as telecommunications), have been nurtured in a world of near-absolute fossil fuel dependency. Indeed, the Internet, a vital global and largely urban network of networks, relies largely on fossil fuel operated hardware, conduit-based webs, and wireless communication systems. If global communications are to be sus-

tained reliably beyond the middle of the 21st century, they must be powered by distributed, ubiquitous, and redundant renewable power supply systems. The required technology arrays with the concomitant policy and action frameworks do in principle exist to bring about manageable postfossil fuel urban environments.

2. HISTORICAL PERSPECTIVES

2.1 The Rise of the Fossil City

Modern cities have grown on a rich fossil fuel supply base, and especially dependent are the largest, most rapidly growing urban agglomerations. The very logic of their global rise and regional spread is founded on the availability of powerful, centralized, and inexpensive fuels—coal, petroleum, and natural gas—yielding fossil urban structures and patterns based on fossil transport, fossil construction machinery, and fossil industrial systems and manufacturing processes. Intensive economies and labor markets are clustered around the centralized and networked city regions, anchored by heavy investments in infrastructure: power, transport, and communications. This has boosted the primacy of cities over—and ultimately detachment from—agrarian hinterlands.

The new cities of the 19th and 20th centuries—and the very cultures they engendered—were a product of the rising combustion economy. The sheer size of London exploded when coal-fired power made this a technological possibility. Historian Lewis Mumford called the phenomenal urban transformations taking place in industrializing countries between 1820 and 1900 “Coketown” or “Paleotechnic Paradise.” Pre- and post—World War I modern city innovations in the Soviet Union, the United States, Europe, and across Asia were brought to life by the electrifying jolt of the new energy technologies. In the 1940s and 1950s, industrial and government-led transport policies revolutionized urban America away from rail-based modes to technologies promoting petroleum, automobile, and tire consumption. The rising car dependency in most cities, combined with the spread of electrified suburban households under expanding power grids, constitutes one of the most vexing challenges of managing urban environments worldwide today.

Fossil fuel-powered industrial transformations that underpinned broad urban structure change were celebrated by the leaders of the great modern design

movements, from the beginning of the 20th century: Italian futurism, constructivism in the early Soviet Union, De Stijl in the Netherlands, the Bauhaus in Germany, the declarations of the International Modern Architecture Congresses (CIAM—Les Congrès Internationaux de l'Architecture Moderne, 1928–1956) and the International Style that spread from the United States throughout the industrialized world.

The fossil machine age is an outgrowth of earlier stages of the Industrial Revolution when water power was a main driver of textile mill operations. Boosted by the industrial application of coal and electricity, it gave rise to Frederick Winslow Taylor (1856–1915) and Henry Ford's (1863–1947) ideas about the mechanization of manufacturing, as noted by Gideon in 1948. In their wake, the increasingly urban and automated global production-consumption systems of the advanced industrial age boosted global power use at an exponential rate.

But industrialization as powered by electricity, coal-fired steam engines, and petroleum combustion motors also meant the rapid growth of cities, driving the search for innovation in urban form. A plethora of utopian premonitions spread as a result, as revolutionary as the technological changes that triggered them. Peter Kropotkin's (1842–1921) and other anarcho-sindicalist influences on ideas about ideal communities were a direct outgrowth of this era; so were Sir Ebenezer Howard's (1850–1928) concepts and the rise of the Garden City movement. The genesis of the Regional Plan Association of New York and its seminal plan of 1929 were of this lineage, as was Frank Lloyd Wright's (1869–1959) Broadacre City (first described in his 1932 work *The Disappearing City*). Ludwig Hilbersheimer's (1885–1967) mass housing concepts and General Motors' Futurama pavilion at the 1933 "Century of Progress" World Exposition in Chicago boldly anticipated the great car-friendly cities to come. Brasilia, Canberra, and all of the British New Towns are fossil technology plans come to life.

The urban environmental phenomenon of the modern suburb followed. Early ex-urban subdivisions sprung up as aspiring new communities along the tracks of electric tramways, such as the famed streetcar suburbs of many North American, European, and Australian cities such as Brookline, Newton, or Jamaica Plain in Boston, or Balmain, Rozelle, or Mosman in Sydney. Soon thereafter the combustion-engine driven vehicle created a new urban reality altogether: the new car suburbs expanding along motorways. Mass media-driven promotion of a new fossil fuel-based lifestyle

spawned the automobile and petrochemical product-based industrial civilization after World War II. This new vision became expressed in a new set of urban form aspirations. Many of these early premonitions proved to be very much practical concepts: already very early in the 20th century the great electric inventions of the telephone and the elevator opened the door to skyscraper cities, first as utopian visions and later as pervasive reality.

Modern carbon-based energy civilization is epitomized in its great landmarks with the seminal work of migrant Bauhaus architects such as Walter Gropius and Mies van der Rohe. Swiss architectural pioneer Le Corbusier's ideas about the radical modernization of cities are widely regarded the most influential of these, boldly calling for the razing of prefossil cities, as exemplified by his *tabula rasa* concepts: the Parisian Plan Voisin; La Ville Radieuse; the famed lone pilot project for a mass housing block, Unite d'Habitation; and in some realized urban projects such as in the Alsatian town of St. Dié. His plan for Chandigarh, the joint capital of the Punjab and Haryana states established in 1953, stands out among the league of pioneering fossil fuel based urban innovations. This simple yet revolutionary doctrine was soon implemented in cities in the United States, the United Kingdom, and countries around the world. It brought urban destruction and blight through wholesale urban renewal projects in what were identified as inner-city slum areas: the neighborhoods of the poor and disenfranchised and the prefossil relics they inhabited.

The postwar years in Europe saw powerful expansion plans realized, taking advantage of the spreading webs of electrification that crisscrossed former farmlands. The Greater London strategy, the European satellite city concepts, and the Stockholm and Copenhagen finger plans gave bold plan form to a wider, worldwide turn to life in cities, driven by structural shifts and opportunities brought about by the rise of the fossil fuel economy and its promises.

Yet the new vision of renewed and more orderly, healthier, and socially more equitable cities did not escape the environmental reality of advanced industrialization. With the burning of fossil fuel arrived sulfur dioxide, nitrous oxide, and a host of other toxic gases as well as carcinogenic airborne combustion particles, creating the smog and air pollution crises of the 1960s and 1970s. With the exception of the greenhouse gases themselves, these have largely been conquered as hazards in the industrialized world while still plaguing large cities of the developing world.

Mounting calls for more densely concentrated and public transport-based forms of urban development since the 1970s were early signs of a declining stage of the fossil-fuel economy when savings in energy conservation and fuel efficiency began to be identified as the cheapest, fastest, and most immediately useful means to reduce emissions. Denser cities were shown to be more fuel-efficient, while car-dependent, low-density urban structures incapable of sustaining public transport have come to be understood as a major hindrance to achieving sustainable urban life.

It is widely understood, however, that efficiency and conservation measures are not sufficient to halt the global rise in CO₂ emissions in the long term. Massive substitutions of conventional energy technologies conveyance with renewable systems of energy supply and use have long been accepted as essential. For these new technologies to be in place in time to be of significant use, they need to be widely introduced as soon as possible. In this sense, transit-oriented development and attempts at re-creating preindustrial urban patterns such as neo-traditional design and other approaches are important interim measures.

Both nature and form of buildings were changed by the development and spread of the revolutionary new fuels and the industrialized manufacturing processes they engendered. Urban environments and architecture changed radically in the new fossil age, breaking with all earlier traditions. The Modern Movement stylistically refined mass applications in electricity, power machinery, air conditioning, industrial steel products, advancing glass technology, mass produced curtain walls, prefabricated building systems, and a number of other highly energy-intensive technologies. The new thinking about buildings, highlighted by the International Style, applied to aesthetically refined, ornament-free, skeletal, industrialized, and largely corporate structures. It became the new global aesthetic of the possible as the advanced fossil age dawned in cities around the world. Buildings became disconnected from their climatic and cultural context due to the end of local resource dependency. An interest in vernacular form and regionalism emerged as a counter-reaction and an attempt to recapture the local language of form: the regional building traditions that had been lost in industrializing change. This movement remained a stylistic artifice since it failed to address the fundamental dynamics of design and development under conditions of unsustainable energy resource practice.

2.2 Local Measures: Fossil Fuel Conservation versus Replacement

As a reaction to these historical perspectives, the building industry has witnessed a certain rise in design responses to regional climatic conditions, as part of a powerful efficiency and conservation push since the 1970s. More recently, the zero-emissions house and office building has become a design concept as part of strategies to introduce renewable energy as an increasing contributor to managing urban energy supplies. In its pure form it is to function without primary fossil or nuclear energy supply input and has been the focus of worldwide research and development efforts for several decades. Also addressed is the question of how to reduce the large amount of fossil energy that is embodied in the building materials themselves, the services contributing to their making, and the energy implications of the very form and design of neighborhoods and cities. The energy household of cities is expressed by the sum total of all goods and services consumed in a given location: hence location-based emissions reduction measures are a primary target of urban energy reform.

The urban transport question, too, has opened up a powerful energy dimension, with two countervailing fundamental positions emerging. The energy conservation strategy points to the massive sprawl and wastefulness engendered by modern car-based cities, with the least dense consuming the highest amount of gasoline per capita. The public transit oriented development and urban consolidation or compact city policies of the 1980s resulted from these observations. The other argument is based on a general conviction that the age of individual transport is inevitably going to stay. It sees the future of personal transport to lie in biofuel and hydrogen combustion or fuel cell-powered vehicles. In this scenario the advocates of denser, more complex urban settings serving cultural and social objectives would lose out, since it paints the prospect of a solar-hydrogen energy revolution that could also accelerate urban sprawl.

2.3 Global Aspects

Global urbanization as a product of the fossil-fuel economy is expressed in the ready availability of inexpensive fossil power for all urban infrastructures, telecommunications, and freight and human transport systems on surface, sea, and air. These bundles of infrastructure at once link cities globally and bypass their regions. As a consequence, globalized urban systems are inherently more vulnerable to the

decline of global fossil fuel supplies than those that rely more on their local and regional human and land resources.

The deployment of renewable energy technologies on a large and especially urban scale has the potential of bringing about a time of differentiated globalization, marked by a distinction between local supplies of food and elementary goods on the one hand and the global trade in services on the other. This condition as brought about by nonfossil production modes would be characterized by a rise of regional economies in support of urban centers, based on regional resources such as productive land for food, biomass, and wind energy production. New ways of reknitting central cities with their regional economies and related spatial structures are already being pursued by a number of communities. These are based on older principles of rural urban support economies, boosting the primary sectors of agriculture and forestry. Cities around the world have started to make concrete links between their renewable energy needs and potential regional resources capable of meeting that need. This movement is also beginning to help spawn new indigenous manufacturing and advanced industry sectors in renewable energy production, supply, and services.

In urban industrial terms, initiatives fall into two categories: new technology development and the accelerated market uptake of applications. Technology push and supply from the 1970s through the 1990s were limited to a number of limited-scale industry efforts and pockets of largely government-sponsored research and development projects. This history is often marked by a mismatch with market demand, especially given the powerful subsidies granted to the fossil energy sector. Yet the international and domestic policy and pricing environment of the early 2000s promises to change fast, heralding massive urban technological and practice changes and a natural integration of technology development and markets.

While a few local urban systems may seem relatively safe from a terminal fossil fuel shock through their reliance on coal, hydroelectric, nuclear or bioenergetic power, no currently utilized alternative energy source alone can keep operational the vast majority of cities. Also, the interconnectedness of the global system makes implausible the survivability of regional pockets of self-sufficiency. The only viable option to secure the continuity of urban civilization in this century is a system-wide turn to a broad portfolio of renewable energy sources based on an overwhelming availability of solar, wind, wave, hydrogen-based, and other renewable forms

of energy generation, combined with massive conservation and efficiency innovations.

3. INNOVATION IN URBAN RENEWABLE ENERGY TECHNOLOGY

Most contemporary fuel sources are due to expire well within the 21st century as widely available and predominantly peacefully contested sources, and much of this reality will become globally pervasive within these next decades. Even conservative industry, national, and international governmental sources estimate that oil will expire by 2050—the depletion of the more easily accessible supply sources is likely to take place already in the 2020–2030 time frame. Given rising use rate scenarios, natural gas is likely to evaporate as a commercially available source by 2040. The oldest and most problematic source of fossil fuel supply, coal, is expected to expire as a globally available source before 2100—provided no large-scale efforts of energy fuel substitution are pursued to stretch its deposits.

Uranium supplies are also estimated to reach depletion levels by the mid-2030s, even if the daunting cost, security, safety, disposal, and public acceptance dilemmas did not weigh so profoundly against their use. Significant growth rates in uranium usage are not anticipated, given the risks involved and a waning willingness among many governments to pursue nuclear fission further. Nuclear fusion, more than half a century after its arrival as a great modern dream, still struggles with enormous conceptual and technical flaws, and the problems associated with huge financial, environmental, and political costs. These include the spectre of broad public resistance to a new generation of nuclear super-reactors.

This outlook is as troubling as it is underemphasized in its public significance. Fossil fuels are the lifeblood of modern economies and the foundation of the petrochemical, transport, and industrial production systems. Hence they also sustain the spatial centers of global civilization: cities and city regions. The paradox is perplexing but can be explained by the daunting magnitude of the problem. While the modern city in its history, form, and growth dynamics is inconceivable without these fuels, a persuasive agenda to deal with the prospects equitably, humanely, and positively still eludes global and local planners alike.

Yet the most optimistic plans about the powering of urban areas describe entire cities as net renewable energy producers. This idea requires a rethinking of

urban-regional alliances as well as an adoption of increasingly firm industry promotion practices. The Australian city of Melbourne, as an example, is in the process of investing in renewable energy producers with the dual aim to reduce its fossil fuel dependency and to promote the development of more advanced industries capable of competing nationally as well as internationally.

In a technologically advanced renewable economy energy supplies no longer exclusively depend on large, centralized supply models but can unfold in a more diverse and differentiated manner, in keeping with the principles of convergence, blurring the conventional distinctions between different technologies and production and consumption. Traditional energy users become net energy generators, for instance, through solar systems or zero-emission, renewable-source based hydrogen fuel cells in vehicles capable of powering homes and external machinery.

Systems convergence dynamics also point to a merger of information technology, telecommunications, and energy systems. While some electric utilities already lease their grids for information transmission purposes, emerging technology goes much further: future energy systems are ubiquitous and pervasive, and networked into virtual power plants. Ubiquitous energy management connotes the hopeful notion that in a renewable-energy based economy a myriad of small and medium-scale providers of energy services could replace the system of large-scale centralized ones. In an urban system this could operate both at the high end described and at low levels of technological sophistication.

Possible high-technology directions contain features derived from information technology and telecommunications mergers, or the blending of these new technologies with energy production and consumption modes residing everywhere, in applications that range from personal equipment to cars and facilities. At the lower end of the scale, in a distributed yet low-cost and low-maintenance environment, small hydropower and solar systems are seen to be capable of leveraging access to the global information network for small remote communities in developing nations.

4. THE URBAN RENEWABLE ENERGY PORTFOLIO

A number of renewable energy technologies and related approaches and systems is available to supplement and even supplant conventional fuel sources in cities. As a main distinguishing feature

when compared to more conventional systems (fossil, nuclear, large hydro, municipal waste incineration), they all vary greatly in scope, style, and systems features, and their economic and physical attributes varies too, including their visual attractiveness. Together, they present an urban power paradigm shift, in the move from the large and passively sourced supply networks to a much wider and dynamic array, or portfolio of systems, many of which are in the control of individual users.

Because the legislative, practical, and financial energy supply structure in a number of countries is not traditionally geared to cope with such simple systems as decentralized small cogeneration plants or grid-feeding domestic solar photovoltaic or solar thermal power plants, significant challenges lie ahead, waiting to be overcome. Fortunately, numerous urban applications are in place, largely in Europe and the United States, ready to be used as best-practice examples.

Buildings, facilities, infrastructures, and even cultural expectations associated with the city are intrinsically linked to and based on the industrial paradigm engendered by fossil fuels. Energy efficiency and conservation have an extraordinarily powerful role to play in reducing the historical dependency of cities on massive fossil fuel input. Ambient temperature stabilizing landscape corridors, native garden-based suburbs, vegetation-covered roofs and walls, natively landscaped atriums, urban agriculture, and better water management all contribute to reducing urban energy input requirements, in this case for cooling.

In supplementation of efficiency and conservation measures, the available power-generation techniques in cities include those shown in the Table I. Many sources listed under Bio/waste are not strictly renewable or environmentally sustainable in their use, especially those relying on waste incineration. They are nevertheless included because many countries do account for them in their renewable, sustainable, new, or alternative energy statistics, as does the International Energy Agency (IEA). Unfortunately, these *de facto* nonrenewable and carbon emitting sources represent the lion's share of many statistically declared sustainable energy components. New, more efficient fossil fuels systems such as advanced coal and gas applications are excluded altogether.

5. CLIMATE STABILITY IN URBAN DEVELOPMENT

Worldwide, municipal policies are marked by a slow but steady growth in commitment to planned urban

TABLE I

Renewable Power-Generation Techniques and Applications Relevant to Urban Systems

Solar	Photovoltaics (PV)	PV panel and array systems	Building/infrastructure
		PV concentrator	Industrial/commercial
		Building-integrated photovoltaics	
	Solar-thermal	Parabolic trough concentrator	
		CLFR technology ^a	
		Parabolic dish/sterling concentrator	
		Power tower/MTSA ^b	
Wind	Turbines	Solar chimney	
		Solar water heater	Buildings/facilities
		Solar pond	Parks
		Large turbine	Grid/region
		Offshore wind farm	
		Small turbines/vertical axis turbine	
Bioenergy	Biofuels/wood	Black liquor	Power stations: Grid
		Forestry crops	
		Forestry residues	
		Timber/paper industrial wastes	
	Biofuel/agro	Agricultural crops	
		Agricultural residues	
		Bagasse	
		Animal waste	
	Biofuel/municipal	Food industrial waste	
		Municipal solid waste	
		Sewage	
		Landfill gas	
Hydropower	Hydro-electricity	Conventional hydropower	Grid/rivers/streams
		Mini and micro hydropower	
Geothermal		Large geothermal	Grid
		Heat pump	Distributed
Ocean		Ocean thermal	Grid
		Tidal power	
		Wave energy	
	Hydrogen	Combustion	Transport
		Fuel cell	Distributed (heat/power/transport)

Assembled by Hiroyuki Akagawa, Technical Research Institute Obayashi Corporation, University of Sydney (2002).

^aCLFR, compact linear solar reflector.

^bMTSA, multitower solar array.

energy practice. As global agreements promise to become firmer under mounting pressures to resolve the greenhouse challenge, and as nations and regional policy units such as the European Union begin to establish clearer greenhouse, energy, and environment frameworks, local institutional pressure mounts to arrive at unprecedented arrangements in urban emissions and energy practice.

As cities are increasingly seen as global emissions performers, distributed energy managers and potential credit trading entities, a new science and

engineering practice is on the rise that interprets cities and city regions as power systems, resource flow units, and point-of-emission locations. The best-known urban environmental and community planning models are based on relative improvement targets and selective means of accounting for greenhouse gas (GHG) emissions, setting the stage for the phasing in of renewable energy systems on an urban scale.

A number of place-based emissions allocation techniques and action approaches are in use. These

include the United States Environmental Protection Agency's support made available to states in compiling GHG inventories, in producing action plans, and in staging demonstration projects. They also include the emissions approximation system developed for the Cities for Climate Protection (CCP) program operated by the International Council for Local Environmental Initiatives (ICLEI). Another accounting approach based on the behavior of large geographical units—one degree of longitude by one degree of latitude—is a program called Global Change in Local Places developed by the Association of American Geographers. There is the potential to apply aspects of the Advanced Local Energy Planning approach developed under the International Energy Agency's (IEA) Building and Community Systems program, or the physical model of the economy developed by the Australian Commonwealth Science, Industry and Research Organization. Another ambitious model is a total-flow, end-consumption, and value-based accounting system advanced at the University of Sydney. It also aims to compare, evaluate, and advance a range of methods.

While most GHG emission sources and mitigation efforts are inherently local, their effects and most easy modes of measurement operate globally. However, the identification of globally diffused GHG levels carries little practical local meaning. For cities to become active and integral participants of any global action program, they need to opt for one or the other GHG accounting or allocation method that reliably links local practice to global aims. A persuasive argument holds that only systems reliably measuring contributions to climate stability are valid and more specifically approaches that embrace a 3.3-ton of a 1990 carbon dioxide-equivalent target by 2050. This target is based on a fundamental equity calculation: that on a per capita basis each person has only an annual 3.3 ton emissions allowance, if oceans and forests are to neutralize excessive carbon emissions. In contrast, Australia and the United States approach 30 tons per annum per person, while most developing countries, including India and China, lie well below this sustainable 3.3 t/p/a level.

An ideal method has not yet been found, and in pursuing it, cities need to apply certain performance criteria. To be informative, universally acceptable and suitable as a basis for coordinated climate stable practice such model would need to be comprehensive, precise, and accurate. To be useful it also has to be pragmatic, affordable, and easily replicable across a large number of communities or geographic units.

Another challenge lies in avoiding a new information poverty barrier. It could develop were local carbon trading to be introduced as based on highly sophisticated and expensive forms of data gathering—poorer nations have less capacity to cope with the technical challenges involved in measuring emissions performance, let alone reliably monitor its progress over time.

The single most widely promoted method is that of ICLEI's CCP program. It is introduced by participating local governments as the first of CCP's five steps to better GHG performance: namely the arrival at an energy and emissions inventory, forecast, reduction target, target achievement plan, and active implementation program to achieve measurable GHG reductions. At these relatively early days of the CCP program measurable success in reaching significant implementation goals has been elusive, while the other milestones have been reached to lesser and varying degrees. It is hoped that in future rounds of development, in association with other methods and given a better understanding of the key barriers to success, a breakthrough in implementation can be achieved. The very imperfections of the method and its omission of significant emission sources makes the CCP system attractive in practical terms but perceived as scientifically ambiguous in globally credible accounting contexts, and of limited efficacy for local improvement monitoring over time, when absolute performance is to be measured.

There are a number of obstacles on the path to a transformed urban energy practice, although none is insurmountable in principle. The most formidable is the fact that local governance structures are not usually geared to end-state or long-range planning, hence the adoption of long-term accounting methods is often resisted. The effective allocation of incentives and accountabilities is difficult too. Notwithstanding the efforts described earlier, there is the persistent question of how to convincingly disentangle resource flows in a local government area not only for carbon accounting but also for scenario modeling purposes, and how to identify agency accountabilities for savings and reduction.

Despite these challenges, the very consideration of broad energy and greenhouse action programs opens up new dimensions for government. It can provide a new development perspective and drive regional realignments, the formation of state initiatives and industrial alliances. It can mean a new sense of empowerment and the opportunity for economic strengthening as well as engender competitive advantages of a more promotional nature.

Governments discover that energy responses can be scaled well to local development paths and growth options. In newly industrializing countries with interest in maintaining considerable levels of autonomy such as China and India, there is a good fit with local traditions and climatic and economic conditions. Local industry can get stimulated while environmental policies find a practical agenda. Comprehensive renewable energy practice on the local level directly reinforces broad quality-of-life aims, as well as healthy water, clean food, and sustained biodiversity levels.

Besides the aim of responsibility devolution in pursuit of global greenhouse objectives, another government reform agenda item is being satisfied as well: better accountability in public service. Outcome-gearred reform and strengthening of local governance in delivering on performance commitments could well be a direct outcome of an indexed and measurable sustainable energy practice. Local consensus, too, is expected to be forthcoming more easily as agendas become more clearly understood and made visible in tangible improvements.

6. URBAN RENEWABLE ENERGY POLICY AND PRACTICE EXAMPLES

In the absence of useful and established patterns of practice, a search is under way for new means of reconciling local government's sectorial concerns, technological opportunities, and shifts in energy markets with global environmental imperatives. There are several hopeful initiatives, focusing squarely on comprehensive local action in urban and regional conversion to a renewable energy supply base.

6.1 Urban Programs: Barcelona's Solar Ordinance

An increasing number of cities search for paths to energy sustainability. One of the most progressive of these is Barcelona. The city has calculated its very favorable solar energy supply potential and in 2000 introduced a regulatory system requiring households and industrial users to provide at least 60% of their hot water requirement utilizing solar systems. The scheme is focused, specific, and practical and based on a precise accounting of domestic energy production and utilization profiles.

6.2 Professional Practice: The European Charter for Solar Energy in Architecture and City Planning

The charter, signed by 30 eminent architects and developed by these under the leadership of German architect Thomas Herzog, was released in 1996, at the Fourth European Conference Solar Energy in Architecture and City Planning, chaired by then minister for construction Klaus Töpfer and parliamentarian Hermann Scheer. The charter was advanced in historical contradistinction to the fossil-age Charter of Athens, referred to earlier. It is comprehensive in attempting to deal with most physical planning challenges in an energy-conscious manner. The charter focuses on the role of planners, the building site, the design and construction process, buildings in use, and the city as a sustainable planning challenge.

6.3 Regional Policy: The Eurosolar Guide to State Politicians—A State Program for Renewable Energy

The renewable energy technology and policy organization Eurosolar has in 2001 issued a blueprint program, for the use by regional politicians in the implementation of renewable energy. It has ten focal areas or considerations: (1) the role model played by state government, (2) the role of state energy agencies, (3) the strengthening of municipal power companies and the placement of energy supply into community hands, (4) research and education, (5) renewable energy in regional development, (6) strengthening of agriculture, (7) the orientation of land-use planning and building toward renewable energy, (8) transport and traffic systems, (9) investment and finance, and (10) state government awards for the exemplary use of renewable energy.

6.4 City Guidance: The World Council for Renewable Energy's Solar City Program

The Solar City approach originally emerged from a new generation of International Energy Agency (IEA) research and development work to pursue citywide applications as integral to the main planning agenda. It is now adopted by the World Council for Renewable Energy (WCRE) as a basic policy and planning basis for cities. It recommends that renewable energy technologies and other means of greenhouse gas

emissions reduction and absorption should be applied in a coherent spatial and social context, as well as within a finite and community-wide time frame.

For application by local institutions, the focus is on the energy supply and technology side, within a comprehensive town planning and design strategy that includes institutional arrangements. The goal is to develop energy technology applications and emissions accounting systems along with performance targets linked to urban development and reform initiatives. Also important are strategies that are based on a consideration of urban-rural linkages and value land use and transport investment choices with a view to their contribution to long-range energy and resource self-sufficiency goals.

The Solar City planning approach presents five areas of focus to cities and city-regions, to be advanced simultaneously. This recommended program constitutes a well-rounded framework for transforming any city's energy regime in a systematic and comprehensive manner.

6.4.1 Solar City Strategies

It is recommended to identify local planning and development approaches that are conducive to the introduction of renewable energy technologies, within a broadly energy-conscious community development approach. To be addressed are strategy, planning tools, organizational arrangements, legislation and standards, incentive structures, public information, and exemplary municipal practice.

By introducing improved ways of adopting renewable energy technologies, a municipal program is to be designed to contribute to climate-stable practice in the building and property development industry, land-use planning, and infrastructure development. It should be framed in ways that help strengthen local governments' efforts to build enlightened community performance and household preferences.

Cities and towns can pursue renewable energy programs through the following methods:

- Direct legislation and standards
- The provision of incentives and disincentives
- Corporate capital asset practice, power purchasing, and pricing
- Institutional reform and improved strategic and general planning practices
- Community action, industry alliances, information, and education

Guidance provided to participating cities is to investigate each of these in detail and develop

advanced means of building improved urban practice approaches.

6.4.2 Setting Targets

The aim is to introduce, enhance, and evaluate suitable approaches that help explain the role of renewable energy technologies in the broader urban energy context. It is important to include both renewable energy technology implementation targets as well as absolute climate-stable carbon dioxide-equivalent emissions measures aiming at a specific future date.

Planning methods based on energy technology introduction and emissions accounting methods may deploy backcasting approaches. This involves the development of alternative urban development growth and technology transformation scenarios, then establishing milestones for technology innovation and emissions to determine workable reduction rates over time.

6.4.3 Urban Renewable Energy Technology, Systems, and Industry Development

The aim is to work with cities in advancing renewable energy technologies and systems and to help promote the renewable energy industry in a way that allows it to serve as model for the rest of the national urban system. The emphasis is therefore on market-led approaches of technology system development and deployment, through pricing, investment, electricity purchasing policies, information, model action, and other means.

It is recommended that optional paths are developed, evaluated, and implemented, suitable for the informed and broad introduction of renewable energy technology portfolios and the use by city governments, municipal utilities, businesses, industries, and households. Special emphasis is to be placed on microgeneration and distributed low-energy production in buildings, facilities, and urban infrastructures. Emerging and potentially competing solar and other renewable energy systems and related urban services are to be assessed for their systematic introduction in ways that are meaningful to cities' development agendas—physical planning, sustainability objectives, organization, services—and their value in targeted emissions reductions.

Outcomes are expected to include technology, systems, and industry development options, suitable for selective and targeted implementation in general and specific action plans—answering the following questions: What can city governments in collaboration with industry and constituent urban

communities do to advance the direct use of renewable energy sources? How can these be useful to residential neighborhoods, industry, and transport? How can they support the generation of electricity in quantity, through solar, wind, biomass, geothermal, and sustainable hydroelectric power? And how can cities as organizations contribute to the development and deployment of industrial and residential application, such as stand-alone power generators, heat pumps, photovoltaics, solar hot water production, or solar cooling?

6.4.4 Understanding Best Practice

The recommendation is to make accessible and apply useful lessons from related initiatives domestically and worldwide. This is to be achieved by studying successful practice in integrated urban energy planning, management, and projects. The activities include an identification of scope and criteria for evaluation, information gathering and documentation, study and evaluation, analysis and description, case study development, and communication and dissemination. The scope encompasses technologies, management practices, and growth strategies. Several main categories of case studies are to be differentiated: comparable cities, urban precincts, and settlement projects but also development policies and programs.

6.4.5 Learning from Action

This important final recommendation is to creatively reflect on program experience derived from the application of such programs. This will help develop a shared understanding of the barriers to community, institutional, industrial, and technological change, with a view toward the planned phasing in of solar and other renewable energy sources on an urban and regional scale.

This structured approach will not only be useful to any city engaged in a renewable energy implementation program, but also be of value in the application of lessons and methods across the national urban system.

SEE ALSO THE FOLLOWING ARTICLES

Biomass for Renewable Energy and Fuels • City Planning and Energy Use • Computer Modeling of

Renewable Power Systems • Public Reaction to Renewable Energy Sources and Systems • Renewable Energy in Europe • Renewable Energy in Southern Africa • Renewable Energy in the United States • Renewable Energy Policies and Barriers • Renewable Energy, Taxonomic Overview • Renewable Portfolio Standard • Urbanization and Energy

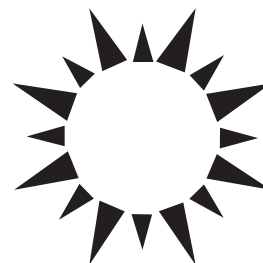
Further Reading

- Byrne, J., Wang, Y.-D., Lee, H., and Kim, J.-D. (1998). An equity- and sustainability-based policy response to global climate change. *Energy Pol.* 26(4), 335–343.
- Calthorpe, P. (1993). "The Next American Metropolis: Ecology, Community, and the American Dream." Princeton Architectural Press, New York.
- Capello, R., Nijkamp, P., and Pepping, P. (1999). "Sustainable cities and energy policies." Springer, Berlin.
- Gideon, S. (1948). "Mechanization Takes Command: A Contribution to Anonymous History." Oxford University Press, New York, NY.
- Girardet, H. (1999). Creating sustainable cities. In "Schumacher Briefings: 2." Green Books, Devon.
- Howard, E. (1902). "Garden Cities of Tomorrow." Swan Sonnenschein, London.
- International Energy Agency (2002). "Renewable Energy Almanac." IEA, Paris.
- Kates, R. W., Mayfield, M. W., Torrie, R. D., and Witcher, B. (1998). Methods for estimating greenhouse gases from local places. *Local Environ.* 3(3).
- Lenzen, M. (1997). Energy and greenhouse gas costs of living for Australia during 1993/4. *Energy* 23(6), 487–586.
- Mumford, L. (1961). "The City in History." Secker and Warburg, London.
- Newman, P. W. G., and Kenworthy, J. R. (1987). "Gasoline Consumption and Cities: A Comparison of U.S. Cities with a Global Survey and Some Implications." Murdoch University, Murdoch, WA.
- Organization for Economic Cooperation and Development (1995). "Urban Energy Handbook: Good Local Practice." OECD Publication and Information Centre, Paris.
- Roulet, C.-A. (2001). Solar energy and the global heat balance of a city. *Solar Energy* 70, 255–261.
- Sartogo, F. (1998). Town planning and the use of renewable energies near Rome. *Renewable Energy* 15, 395–400.
- Scheer, H. (2002). "The Solar Economy. Renewable Energy for a Sustainable Global Future." Earthscan, London.
- Steemers, K. (2003). Energy and the city: Density, buildings and transport. *Energy Bldgs.* 35, 3–14.
- Waals, J. F. M. v. d. (2000). The compact city and the environment: A review. *Tijdschrift voor Economische en Sociale Geografie* 91, 111–121.

Renewable Energy in Europe

JOACHIM NITSCH, WOLFRAM KREWITT, and
OLE LANGNISS

German Aerospace Center
Stuttgart, Germany



-
1. Introduction
 2. Present Use of Renewable Energy Sources in Europe
 3. Policies to Foster Renewable Energies in Europe
 4. Potentials, Costs, and Environmental Characteristics of Representative Technologies
 5. Prospects of RESs in Europe

Glossary

biomass All organic substances that can be used for energy production, including industrial, commercial, and agricultural wood and plant residues; municipal organic waste; animal manure; and crops directly produced for energy purposes. Biomass is available as solids (wood, straw), liquids (biofuels), and gases (biogases).

EU-15 The 15-member European Union countries (Western Europe is used as a synonym for the EU-15 extended by Switzerland, Norway, and Iceland).

final energy The energy used by consumers in final form: electricity, heating fuels, and transportation fuels.

geothermal energy Natural underground heat used for heating purposes either directly or by means of heat pumps (near-surface heat and hot underground water) and for electricity generation (heat extraction from hot rock).

German Renewable Energy Resources Act An act passed in March 2000 that obliges grid operators to buy electricity generated from renewable energy sources at a given price (feed-in model). The price, which depends on the technology and the year the power-generating installations were built, decreases for installations built subsequent to implementation of the act. The overall costs are passed on equally to all final consumers of the electricity.

greenhouse gases Airborne gases contributing to climate change, including water vapor, carbon dioxide, methane, nitrous oxide, and synthetic, volatile chemicals (e.g., tetrachloro- and tetrafluoromethane, sulfur hexafluoride, and hydrofluoroalkanes). The most im-

portant contribution to climate change comes from carbon dioxide.

primary energy The energy directly embodied in natural resources, including fossil fuels (coal, lignite, crude oil, and natural gas), nuclear energy (in the form of heat produced during fission), and renewable energy sources.

renewable energy sources All natural energy flows that are inexhaustible (i.e., renewable) from an anthropogenic point of view: solar radiation; hydropower; wind, geothermal, wave, and tidal energy; and biomass.

1. INTRODUCTION

There are various and somewhat complementary reasons to foster the growth of renewable energy sources (RESs) in Europe. A major incentive for RES policies in the past two decades has been to reduce the environmental impact of energy use both locally (e.g., pollutant emission reduction) and globally (e.g., greenhouse gas reduction). In some countries, concerns about the safety of nuclear power generation have motivated the search for RESs. Another motivation for replacing foreign fossil and nuclear fuels with domestic RESs relates to security issues and Europe's growing dependency on foreign energy sources. In 1999, half of the energy demand of countries in the 15-member European Union (EU-15) was met by imported energy, and this share is expected to increase to 70% within the next two to three decades. In particular, the dependence on oil could reach 90% by 2020. Further motivating studies on developing RESs are the social and economic benefits, such as the creation of jobs, the creation of new industries with good export prospects, and the positive structural effects on regional economies. The characteristics of RES development lead to public benefits in all of these areas. Additionally, the growing integration of Europe

continues to highlight the importance of the future development of RESs.

Currently, individual national goals and political measures substantially determine the actual growth of RES industries. In addition to national goals for RES development, the EU Parliament and the EU Commission have defined broader goals for the contribution of RESs by 2010. Other European long-term energy scenarios, from 2030 to 2050, also deal with RESs, and some call for future energy systems based on significantly high contributions from RESs. In addition, the EU affects the promotion of national RES programs within a legal framework. The Commission is anxious to coordinate national promotion efforts, which currently vary greatly.

Markets for RESs are increasingly being established throughout Europe; the dynamically growing wind power industry has been particularly active in this regard. It is foreseeable that a uniform strategy will be developed within this decade for further promoting and implementing mid- and long-term goals for RES technologies in Europe. This process, which was an essential point on the agenda of the Johannesburg World Summit for Sustainable Development in 2002, is presently supported by a growing awareness of the importance of increasing the development of RES technology worldwide. In summary, there is every indication that Western Europe will strengthen its position as the world's leading region in utilization and commercialization of RESs in the next few years.

2. PRESENT USE OF RENEWABLE ENERGY SOURCES IN EUROPE

2.1 Energy Sources

In Europe, modern RES technologies were explored thoroughly for the first time after the oil embargo/price crisis of 1973. Notably, market introduction of renewable energy technology started in about 1985, but the RES industry has become vital only during the past decade. Accordingly, relevant statistical RES data have been systematically compiled only over the past decade, although reliable and consistent statistical RES data, collected since 1989, exist for all 15 countries of the European Union and for Western Europe (defined here as the EU-15 plus Switzerland, Norway, and Iceland). Earlier data from various European countries are fragmentary, inconsistent, and less reliable. Most energy source data are available up to 2001; the country-specific data have

been compiled up to 1999/2000. Only a few eastern European countries have provided RES data, and their RES contributions today are still very small.

Of renewable energy sources presently utilized in Europe, as well as in other continents, two dominate: hydropower and firewood (for traditional heating purposes). Both resources were the main basis of human energy supply before industrialization, and even with the emergence of fossil and nuclear energies, both have retained importance. In 1970, hydropower met 30% of European electricity demand; by 2000, this had decreased to 17%. The current degree of exploitation of hydropower and firewood today is considered high, thus the contribution of these two energy sources cannot be increased considerably in the future. Primary energy production from RESs in Europe has risen continuously since 1990 (Fig. 1). Within the past decade, production increased by 40%, accounting for 6.4% of total primary energy production in 2001. The current RES production of 3850 PJ/year (a petajoule is 1000 trillion, or 10^{15} , joules) is provided mainly by biomass (62%) and by hydropower (31%). Geothermal energy (4%) and wind energy (3%) contribute far less. The usable final energy production from RESs amounts to 3350 PJ/year, corresponding to 8.5% of the total final energy consumption in the EU-15 countries. Vehicle fuels contribute only 30 PJ/year; usable RES heat (1880 PJ/year) meets 9.5% of total heat demand, and RES electricity (1460 PJ/year $\approx$ 406 TWh/year; 1 terawatt-hour = 1×10^{15} kWh) comprises 15% of total electricity generation (Fig. 2). About 80% of this electricity comes from hydropower. The use of biomass (10%) occurs predominantly in wood-fueled power plants (60%), although biogas and sewage gas (10%) and municipal waste (30%) also contribute significantly. The contribution of wind energy has grown considerably, to 6.6%, and geothermal energy contributes 1.2%; solar electricity presently contributes an insignificant amount (0.1%). The dominance of biomass in heat production is notable. Individual firewood heating contributes about 60%, and another 20% comes from wood- and waste-fired district heating systems; industrial wood-fired heating systems contribute 18% and biogas contributes 2%. Solar and geothermal energy contribute only small amounts to heating (about 1% each).

2.2 Use Different Countries

Renewable energy sources contribute to the energy supply in individual EU member states in quite

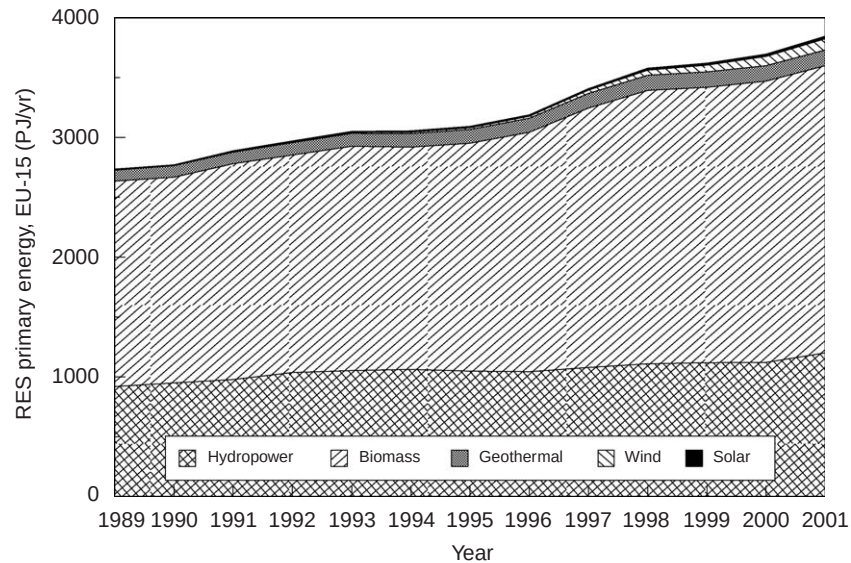


FIGURE 1 Primary energy production using different renewable energy sources since 1989 in the member countries of the European Union.

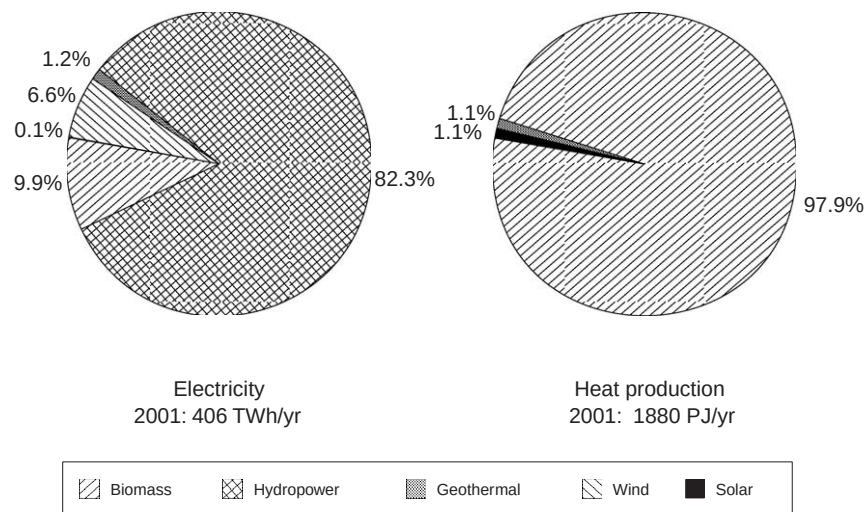


FIGURE 2 Contributions of renewable energy sources to final energy production in the EU-15 for the year 2001.

different proportions (Figs. 3 and 4). This is to be expected because the actual RES contribution is determined by a number of different parameters:

- Country-specific potential for each type of RES.
- Existence of other low-cost and abundant domestic energy resources (e.g., coal, natural gas).
- Structural boundary conditions, such as population density, structure of settlements, ratio of rural to urban areas, extension of electricity grids, and district heating networks.

- Economic conditions, such as national energy prices and taxes.
- Level of economic wealth and of total and per capita energy consumption.
- Conditions of energy policy, such as incentives and instruments to promote RES technology and national energy goals.

Renewable energy resources contribute to meeting a relatively high proportion of energy requirements in countries with large land areas, correspondingly low population density, and abundant hydropower.

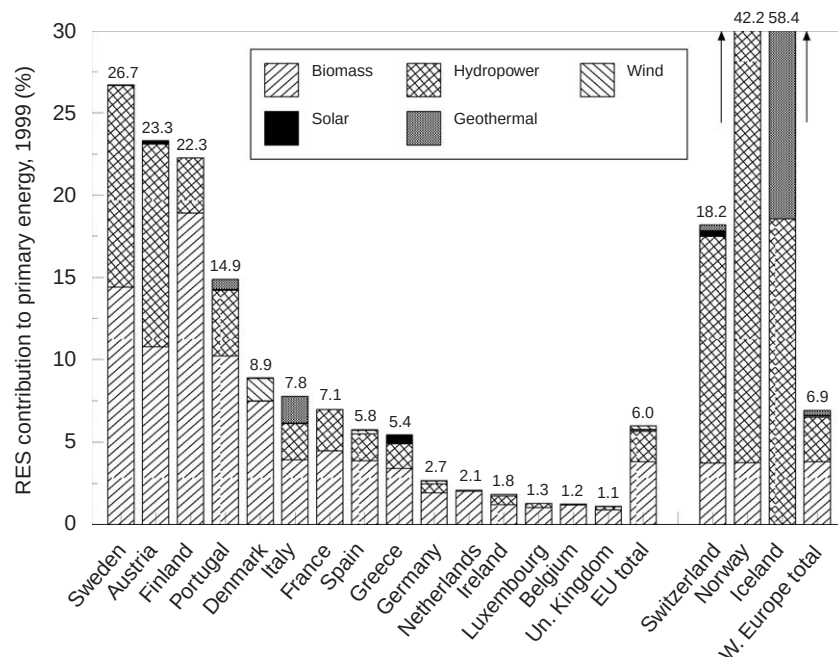


FIGURE 3 Contributions of renewable energy sources to primary energy consumption in the EU member countries and Switzerland, Norway, and Iceland in 1999.

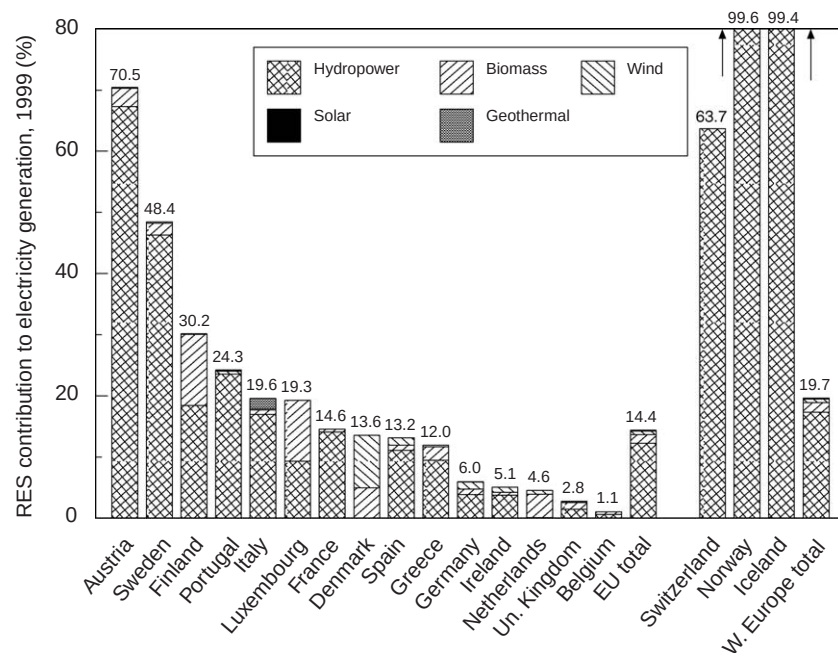


FIGURE 4 Contributions of renewable energy sources to electricity generation in the EU member countries and Switzerland, Norway, and Iceland in 1999.

Sweden, Finland, and Austria fit this profile, with RESs contributing more than 20% to primary energy consumption and electricity generation. In densely populated countries, RESs contribute less than

10%—2.7% in Germany, 2.1% in The Netherlands, and only 1.1% in the United Kingdom. On the other hand, there are strong disparities among countries with similar RES potentials and comparable

infrastructures. For example, in densely populated Denmark, RESs contribute 9% to primary energy consumption and 14% to electricity generation, showing the importance of well-defined political goals and ambitious promotion measures. Such goals and measures, however, primarily affect development of the “new” RES technologies (wind energy in Denmark). RES market development has seen little progress in the United Kingdom.

The contribution of hydropower and geothermal energy increases considerably with the inclusion of Switzerland, Norway, and Iceland, which are geographically, economically, and culturally part of Western Europe (Figs. 3 and 4). RESs in Norway and Iceland provide about 50% of the primary consumption and 100% of the electricity; in Switzerland, these values are 18 and 64%, respectively. Thus RESs accounted for 7% of primary energy consumption and 20% of electricity generation in 1999 in Western Europe. The estimated increase for 2001 is 0.5%. The five countries with the highest amount of RES primary energy are France, Italy, Sweden, Norway, and Germany, which together account for 62% of total RES primary energy production. The highest amount of RES electricity is generated in Norway, followed by France, Sweden, Italy, and Austria; these five countries account for 68% of total RES electricity production.

2.3 Technologies

The development of the “new” wind and solar technologies is of great importance for the future contribution of RESs to energy supply. Although the present wind and solar technology contribution of 0.4% to total primary energy consumption is still very small, the growth of these industries has been considerable in the past 6 to 8 years. Today they provide 10 times the energy of 10 years ago. Wind energy shows the most remarkable growth dynamics; its contribution is now reaching energetically relevant dimensions (Fig. 5). Since 1992, the average growth rate has been 35%/year. At the end of 2001, the installed capacity in Western Europe reached 17,250 MW, about 70% of the total global wind power; 1% of the Western European electricity demand is generated in this way. With 8750 MW, Germany alone accounts for 50% of the installed capacity; Spain, with 3300 MW, follows and is rapidly increasing wind installations. Spain has surpassed the pioneering wind country, Denmark, where growth decreased strongly because of a recent change in energy policy priorities. Yet, with 9%,

Denmark still has the highest share of wind power for total electricity generation.

Photovoltaics also showed similar high growth rates in the past 8 years (Fig. 6), but with an installed power of 300 MW in Western Europe (280 MW in EU-15), the contribution of photovoltaics to electricity generation (0.01%) is still negligible. Even more noticeable than in the case of wind energy, the growth rate of photovoltaics is dominated by development in Germany, which accounts for 60% of the total installed power in Europe. This is the result of the very favorable conditions set up within the framework of the German Renewable Energy Resources Act and the additional “100,000 PV-installations” program. Considerable growth in photovoltaics has also occurred in The Netherlands, which has reached

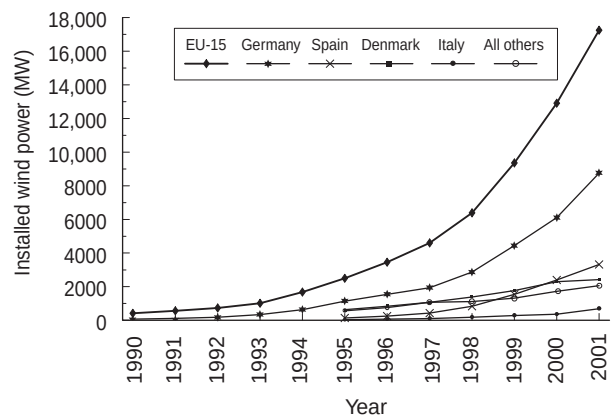


FIGURE 5 Growth of installed wind power output in selected EU member countries and the EU-15 from 1990 to 2001.

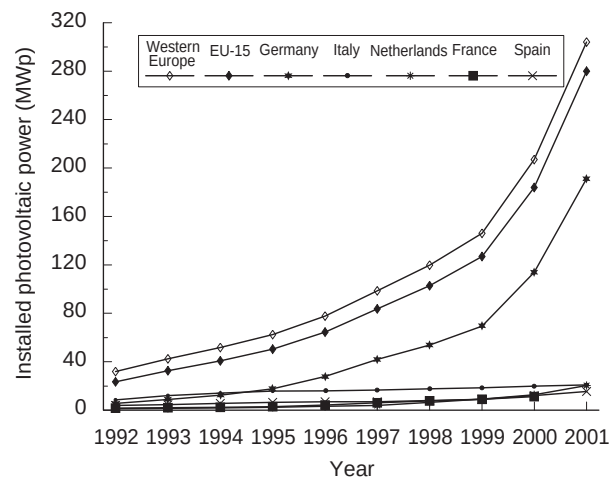


FIGURE 6 Growth of installed photovoltaic power output in selected EU member countries, the EU-15, and Western Europe from 1992 to 2001.

the power level attained by Italy (20 MW). Nevertheless, related to population, Switzerland has the highest density of photovoltaic installations, with 2 W per capita.

There has also been a solid increase in installed solar-thermal collector area, but growth has not occurred as rapidly as has growth of wind power or photovoltaics. In Western Europe, an area of $12.8 \times 10^6 \text{ m}^2$ (EU-15 = $11.9 \times 10^6 \text{ m}^2$) supplies 0.01% of the total heat demand (Fig. 7). In Germany, Austria, and Switzerland, where climatic conditions are far less favorable than in southern European countries, the installed solar-thermal collector area is surprisingly high, primarily due to favorable incentives and promotions. In addition, nature conservation and resource protection rank much higher in public opinion in these countries, driving favorable development policy. The reverse of this, public disinterest, may be the reason that, excluding Greece, the contribution of solar-thermal collectors is currently very low in southern European countries, despite the high technical potential. In Greece, solar collectors for water heating are cost competitive against electrical water heaters.

A discussion of promising RES technologies is not complete without mentioning improved biomass conversion technologies such as advanced gasification combined with internal combustion engines, Stirling engines, or fuel cells. These technologies have future potential to convert the abundant supply of biomass more efficiently into electricity and heat than is the case today. Also, improved technologies for digestion of agricultural and municipal waste residues will be of growing importance.

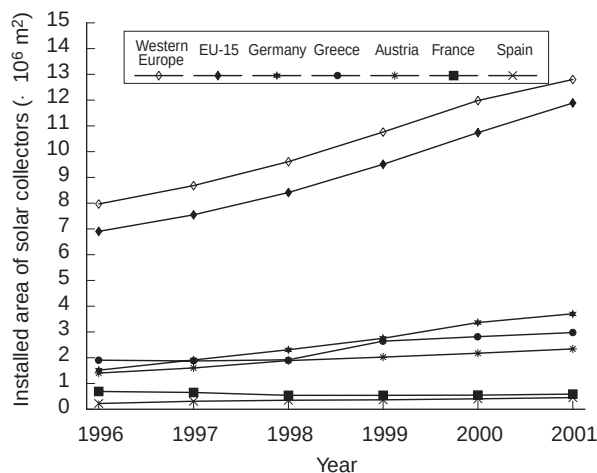


FIGURE 7 Growth of solar-thermal collector area in selected EU member countries, the EU-15, and Western Europe.

3. POLICIES TO FOSTER RENEWABLE ENERGIES IN EUROPE

An important guideline for a well-coordinated RES policy strategy involves setting definite targets, within specific time frames, for RESs to share in meeting the needs of total energy supply. Certainty in policy is a positive signal to potential investors, and attracting private investments is particularly crucial for RES; an accelerated deployment of RES technologies requires considerable funds, usually in excess of funding specified in public budgets. Different approaches are needed to support deployment of RES programs. A comprehensive strategy with a variety of different instruments, rather than a single instrument, is required to address RES development for a range of applications. Moreover, RES policies must address noneconomic barriers, research and development needs, and the cost gap between RES providers and conventional energy providers. Overcoming the renewable/conventional cost gap is a central building block of any RES policy. For this purpose, energy or emission taxes on conventional energy providers can be raised, reflecting the external costs of such fuels. The costs of a RES to the user can be reduced by providing grants, soft loans, or tax benefits. The cost gap can also be overcome by implementing the renewable portfolio standard, i.e., creating a demand for RESs through obliging electricity suppliers to generate a certain percentage of power from renewable sources and offering homeowners and businesses incentives to purchase renewable power. The voluntary purchase of "green" electricity (electricity from RESs), although widely discussed, has not met expectations in Europe so far, and can only supplement, but not substitute for, public support policies.

Establishing RESs via political and regulatory structures is a further important element for a successful RES policy. The present energy supply system and the political and regulatory framework conditions have evolved in a mutual adaptation process. Certain characteristics of RESs, such as their decentralized, dispersed nature, are disadvantaged in many cases by the present regulatory framework. A successful RES policy needs to address such noneconomic barriers for RES. This includes fair access to the electricity grid, adapted building permit procedures for wind power plants, and consideration of solar-thermal collectors in building codes. Workmen, tradesmen, and engineers need to be educated comprehensively about RES technologies and their applications. Schools and universities

need to incorporate RES studies in their curricula. Last but not least, there is an ongoing demand for research and development to exploit the cost reduction potential through technical progress.

Energy policy happens on different levels within the EU, with its institutions, the European Parliament, the European Commission, and the European Council. The promotion of renewable energy sources at every level of the European Union is gaining more and more importance. On the one hand, the general framework conditions for European energy markets are very much targeted toward creating a common market for energy, with equal conditions for all market players across the EU. As an example, the EU issued a directive, "Common Rules for the Internal Market in Electricity," which has created fairer access to the electricity grid for independent power producers using RESs and green electricity suppliers. The European Community Guidelines on State Aid for Environmental Protection allow member states to grant operating support to new RES power plants of up to 5¢/kWh, referring to external costs of conventionally fueled power plants. The European Union is also an active player in promoting RESs directly. In 1997, the European Commission issued a White Paper, "Energy for the Future: Renewable Sources of Energy," setting a goal of doubling the share of RESs for primary energy supply by 2010 and describing scenarios and policy strategies to reach this goal. As a follow-up, a "Campaign to Take-Off" was launched in 1998. The EU's own financial means allocated to this campaign have been rather limited, with \$74 million (U.S. dollars) allocated over 5 years compared to the total required investment of \$20 billion (U.S.). Another European directive ("On the Promotion of Electricity Produced from RES in the Internal Electricity Market") has set goals for the EU member states regarding the share of RESs for electricity production by 2010. Although the directive failed to establish a common European instrument to foster RESs for electricity generation, it has created some momentum to establish support policies on the national level, as seen recently in the United Kingdom and in Sweden. A political review of all national RES policies is scheduled for the end of 2005, with the goal of creating a basis for a common European support instrument for electricity from RESs.

National support mechanisms remain the most important means to foster development of RES technologies. For example, national minimum price standards (also referred to as fixed feed-in tariffs) have, in particular, promoted wind power in Denmark, Germany, and Spain. Favorable conditions

have been created for biomass-fueled district heating in Sweden by imposing high taxes on conventional energy providers and a via CO₂ tax refund. Soft loans and direct investment grants have caused the demand for solar collectors to soar in Germany. Examples of regional and municipal RES policies include the solar ordinance in Barcelona, requiring real estate developers to install solar water heaters, and the green power purchase policy in some Dutch municipalities.

Generally, RES policies have been focused on the electricity sector rather than on transport or heating, mainly because there is a long tradition of state intervention in the electricity sector. Minimum price standards, bidding schemes, and renewable portfolio standards, also referred to as RES quotas or green certificates, have been the major way to support electricity generation from RESs on the national level. Minimum price standards require the grid operator or the default electricity supplier to purchase electricity from RES generators at fixed premium prices. It should be noted that a minimum price standard not only regulates the price, but also regulates grid access and power purchase. Within bidding schemes, RES capacity is publicly tendered periodically and power purchase contracts are awarded to the winning bids. With renewable portfolio standards, electricity suppliers are obliged to generate a certain share of their electricity supply with RESs. The engaged parties comply with the obligation by presenting tradable "green certificates," certifying the generation of a certain amount of electricity. Thus, these certificates have an economic value, generating extra income for RES electricity producers.

Countries with minimum price standards (e.g., Germany since 1991, France since 2001, Spain since 2000, and Denmark until 2000) have seen the largest growth of RES electricity. This particularly applies to wind power. At the same time, a viable RES manufacturing industry has been established in these countries. To organize political support and create local acceptance, it has been proved successful to spread ownership among many, preferably local, people. Even though it is not appropriate to attribute the success in RES deployment solely to a single policy instrument, it is clear that a well-designed minimum price standard together with supplementing policies (such as simplified building permit procedures) seems to be the most effective way to support the introduction of RES electricity. Nevertheless, as indicated by experiences in different countries, proper design of a very specific

support policy is even more crucial than the type of policy.

England and Wales introduced a bidding scheme called Non-Fossil-Fuel Obligation (NFFO) in 1990. In five bidding rounds, between 1990 and 1998, developers of RES plants could bid in different technology slots (e.g., wind power, waste to power, hydropower). The winners with the lowest bid generation costs were awarded with a 15-year power purchase agreement. The bid prices ranged between 45% (hydropower) and 70% (wind power) between the first and the last bidding rounds. Due to different conditions in the bidding procedures and the awarded power purchase agreements, the bids were not directly comparable to each other. Moreover, the large wind project put forward in the last bid round in 1998 has not been commissioned at the low average bid price of 4.5¢/kWh. Apparently, the bid prices were not economically feasible. Although the United Kingdom has the largest wind power potential in Europe, the NFFO failed to foster wind-generated and other RES electricity, perhaps due to inherent weaknesses of the NFFO and also to a lack of appropriate building permit procedures. The wind project has been abandoned.

Renewable portfolio standards have been discussed widely and have been introduced in Austria, Belgium, Italy, Sweden, and the United Kingdom. Although such mechanisms are promising in theory, practical experience has been limited and rather mixed in Europe. Larger providers are more likely than small generators to take the risk of selling electricity and certificates under uncertain conditions. Instead of a wide range of different RES technologies, only the most cost-effective technology will likely be supported at a given time. Long-term

contracts rather than spot markets will govern transactions between RES providers and the obliged parties, thus undermining competition. Actually, the different designs of the national renewable portfolio standards hinder rather than enable the free trade of certificates between different countries.

Increasing prices for conventional fuels have been an effective method to deploy biomass and other RES technologies in the heating sector in Northern Europe. Widespread district heating grids support the application of RES technologies in these countries. Household solar collectors have been successfully promoted by tax benefits and direct investment grants. Building regulations allowing only a certain maximum fossil-based heat demand for new building are another effective way to promote RESs. Austria has been extremely successful in deploying solar collectors via grass-roots do-it-yourself construction groups. In some countries, the RES use in the transportation sector is supported by exempting biomass-based car fuels from taxes; this has led to a growth of demand in Germany since 2001.

An overview of promotional systems for RES technology and use in the EU-15 is given in Table I. It is apparent that most countries are using either the feed-in tariff model (minimum price standards) or the certificate trading model (the quota model). Bidding schemes, originally introduced in the United Kingdom, are used only in Ireland. The feed-in tariff model has turned out to be the most successful instrument in terms of installed RES capacity, but an increasing number of countries are considering the certificate trading model as the future winner. Possibly a mixture of both will be used in the future, because green certificates also can be combined with feed-in models.

TABLE I
Overview of Promotional Instruments for RES in EU-15 Countries (2002)

Instrument ^a	Country ^b														
	AU	BE	DK	FI	FR	GE	GR	IR	IT	LU	NL	PO	SP	SW	UK
FIT	X		(X)		X	X	X			X		X	X		
BID								X							
SUB			(X)	X		(X)	(X)	(X)		(X)	X			(X)	
CTM	(X)	X	Xp						X		(X)			Xp	X

^a FIT, feed-in tariffs; BID, bidding system; SUB, subsidies, tax reliefs; CTM, certificate trading model; X, main instrument; (X), additional instrument or combination with main instrument; p, proposed.

^b AU, Austria; BE, Belgium; DK, Denmark; FI, Finland; FR, France; GE, Germany; GR, Greece; IR, Ireland; IT, Italy; LU, Luxembourg; NL, The Netherlands; PO, Portugal; SP, Spain; SW, Sweden; UK, United Kingdom.

4. POTENTIALS, COSTS, AND ENVIRONMENTAL CHARACTERISTICS OF REPRESENTATIVE TECHNOLOGIES

4.1 Technical Potentials

To provide energy for mankind, ambient energy, in terms of annual increases in biomass and the radiation, wind, water, and geothermal flows, has to be transformed into usable forms of energy such as electricity, process heat, or fuels by means of technical installations and processes. Determination of the potential energy and productive capacity of an energy technology requires consideration of a great number of physical, technical, and structural restrictions. To achieve a realistic, practical picture of the possible applications of a renewable energy source, it is necessary to add further boundary conditions, mainly economic and ecologic. In view of the huge amounts of ambient energy flows—e.g., the solar radiation reaching Earth's surface during 3 hours corresponds physically to the total annual global energy consumption—to meet future energy demands, it is sufficient to produce energy carriers only in per million shares (in maximum percent shares).

Numerous analyses of current RES technologies provide a detailed and sophisticated insight into the extent and structure of the technical potentials of RESs. Table II presents these data for Western Europe. The values are explicitly downsized according to infrastructure specifications and ecological restrictions. The comments in the table point out the most important assumptions in this respect. It can be assumed, therefore, that these potentials can be tapped completely as soon as they become economically viable. Compared with present energy costs, the majority of these technical potentials cannot be exploited without public subsidies. The total RES potential for electricity generation is almost double the present electricity generation of Western Europe and the total RES potential for heat production meets 75% of the present heat demand. Altogether, based on these calculated RES technical potentials, 62% of the present primary energy consumption could be met. The different RES technologies are exploited very unequally. Hydropower potential is exploited to a large extent, if the construction of new plants on nonregulated rivers is excluded. Biomass residue utilization also already takes place to a large extent. All other RES technologies are currently exploited to a very small degree. The potentials of

RESs are 10 times the present amount of energy that is provided from RESs in Western Europe.

In the long-term, if markets for RES technologies are established, far more renewable energy sources can be made accessible. The following examples of potential market development are from the point of view of Western Europe:

- An extended application of offshore wind converters on European coasts, with an additional electricity generation potential of about 2000 TWh/year.
- Cultivation of energy crops on unused agricultural areas in connection with the planned enlargement of the EU to Eastern European countries, adding an additional 30 million ha (i.e., 3500 PJ/year primary energy equivalent).
- Further exploitation of geothermal resources, yielding additional electricity generation of up to 1700 TWh/year.
- Import of electricity from solar-thermal power plants in North Africa to Western European countries by means of an European–Mediterranean electricity grid, allowing, in principle, the additional mobilization of some 10,000 TWh/year of electricity.

It is obvious that these additional RES potentials are available mainly as electricity. There is no sense in further exploiting RES heat potentials, because heat supply at low temperature (up to 150°C) has to be adjusted to the demand for space heating, warm water, and low-temperature process heat. In contrast, surplus RES electricity can be transformed by electrolysis to hydrogen, which can be used as a fuel for transportation as well as for high-temperature process heat in industry. The above-mentioned additional technical RES potentials add up to more than 80,000 PJ/year of primary energy. Thus, even an increasing energy demand in Western Europe can, in principle, be met in the long term solely by RESs.

4.1.1 Concentrating Solar-Thermal Power Plants: A Representative Innovative RES Technology with Large Potential

Concentrating solar-thermal power plants (CSPs) are able to supply very large amounts of electricity and process heat in the earth's sunbelt. Like any conventional power plant, their integration into the grid does not require new measures for backup capacities. On the contrary, CSPs allow for a smooth transition from today's fossil-based power technology to a future energy economy based essentially on renewables. The history of the solar electricity generating

TABLE II

Technical Potentials of Renewable Energy Sources in Western Europe

Renewable source	Electricity generation (TWh/yr) ^a	Heat production (PJ/yr) ^b	Primary energy (PJ/yr) ^b	Used in 2001 (%)	Comments
Hydropower	640		2270	79	New plants only on regulated rivers; includes reactivation and modernization of existing plants
Wind energy	1580		5680	1.8	Restricted potentials onshore (630 TWh/yr); offshore = 950 TWh/yr; acceptance is high and ecological interaction is as low as possible
Photovoltaics	585		2105	0.03	The assumed 2300-km ² module area = 25% of suitable roofs and 100% of suitable facades (1000 km ²)
Biomass residues	185	2350	3350	52	Residues from forestry, agriculture, and the organic fraction of municipal waste; for energy crops, the present excess agricultural area in the EU-15 (9 million ha) is assumed; 50% use in cogeneration plants, 50% direct heat production, no fuel production assumed
Energy crops and short-rotation forestry	85	1050	1500		
Thermal collectors		6570	7300	0.2	The assumed 4000-km ² collector area = 75% of suitable roofs; contains also large collector fields with seasonal heat storage
Geothermal energy	350	5000	9155	1.4	Only known resources (lowest value of a variety of estimates)
Solar thermal power plants	2000		7200	0	Suitable sites for 100 to 200-MW solar thermal power plants in Southern Europe (Spain, Italy, and Greece, including Mediterranean islands)
Total RES potential	5415	14,970	38,560	12	
Consumption, 2001	2800	20,000	62,500		
Share of RES potential (%)	193	75	62		

^aTWh, Terawatt-hour; 1 TWh = 1×10^{15} kWh.^bPJ, Petajoule; 1 PJ = 10^{15} J.

systems (SEGs) in California shows impressive results. Steam cycle plants with up to 80 MW capacity using parabolic trough collectors have been in commercial operation for more than 15 years.

Nine plants with a total of 354 MW of installed power are feeding the California electric grid with 800 GWh/year. The plants have a proved maximum efficiency of 21% for the conversion of direct solar

radiation into grid electricity. Considerable cost reductions have been achieved, with electricity costs ranging today between 10 and 15¢/kWh in solar-only mode. Advanced technologies, mass production, economies of scale, and improved operation will allow reductions in solar electricity costs to a competitive level within the next 10 to 15 years, or as soon as a worldwide capacity of 5000 MW is achieved. Hybrid solar and fossil fuel plants, making use of special financial schemes, can already deliver competitively priced electricity in favorable locations.

In sunbelt countries, CSP is able to reduce the consumption of fossil energy resources and the need for energy imports, enabling diversification with a resource that is very accessible. Moreover, process heat from combined generation can be used for seawater desalination, helping to address the challenge of a growing water scarcity in many arid regions. In the medium term, large electricity grids such as the Euro-Mediterranean Power Pool via high-voltage direct current transmission will allow for intercontinental transport of RES electricity (Fig. 8), thus connecting sites with large renewable electricity resources in Europe (wind, hydro, geothermal) and North Africa (solar). This concept may be of outstanding importance for the energy future of Europeans, because it can help to stabilize political and economic relations between the countries of an enlarging European Union, as well as between the EU and the North African countries. Furthermore, it can help by contributing to a considerable reduction of the growing dependence on Middle East oil and gas resources. Primarily, CSP, in combination with other powerful RES technologies, will effectively reduce the risk of conflicts related to energy and water scarcity and to climate change.

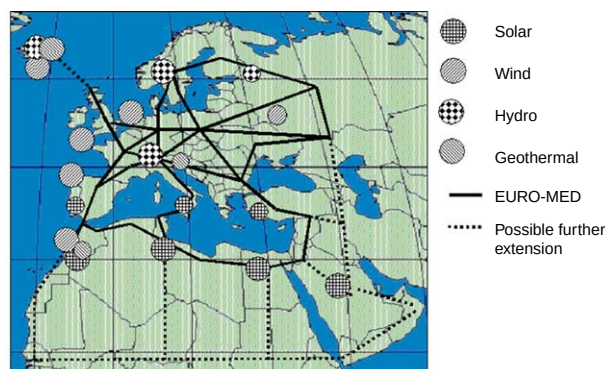


FIGURE 8 Planned extension of European high-voltage electricity grids to a European–Mediterranean (EURO-MED) pool and possible further extension in order to use renewable energy sources widely.

4.2 Costs of RES Technologies

In order to evaluate the chances of a significant commercial RES utilization in the medium term, the costs of energy production, compared to conventional energy supply, are of outstanding importance. The outcome defines which of the RES technologies are economically exploitable under the existing circumstances, and what measures are required to achieve cost competitiveness in a given time frame. It can be assumed that the cost differences between RESs and conventional energies will decrease because of further technical development and growing production rates of RES technologies, and because of an increase in prices of conventional energy sources in the medium term. Thus, a growing share of the technical RES potentials can be exploited economically in the long term. By promoting investments in RESs for a certain time or by influencing the price of energy through taxes or comparable instruments, this process can be accelerated. If future shortages of oil and natural gas or CO₂ sequestration from coal will substantially raise energy price levels, RES providers can offer even lower prices, compared to energy from fossil sources.

A survey of energy generation costs of typical RES technologies in Western European countries, representing the present status and the likely achievable cost reductions in the future, shows a very large spread (Figs. 9 and 10). This results from the great variety of energy sources and their regional differences, the broad spectrum of conversion technologies, the very large range of capacity (from about 1 kW to some 100 MW), as well as different prices for biofuels and for credits in the case of cogeneration plants.

Excluding photovoltaics, for which actual costs are typically 70¢/kWh in Central Europe and 40¢/kWh in Southern Europe, the representative cost of RESs under Western European conditions ranges from 2.5 to 15¢/kWh (Fig. 9). Only a few RES technologies generate competitive electricity today, e.g., older depreciated hydropower plants or cogeneration plants using cheap biomass residues. For that reason, these sources are already widely exploited. A large percentage of RES technologies offer electricity in the range of 7–8¢/kWh, including wind plants in coastal regions, hybrid solar-thermal power plants, and biogas and biomass gasification plants in cogeneration mode. Hence, these RES technologies require further financial subsidies for some time. If market growth of RESs takes place on a large scale, the electricity generation costs of most RES technologies will decline to 4–5¢/kWh. The newest fossil power

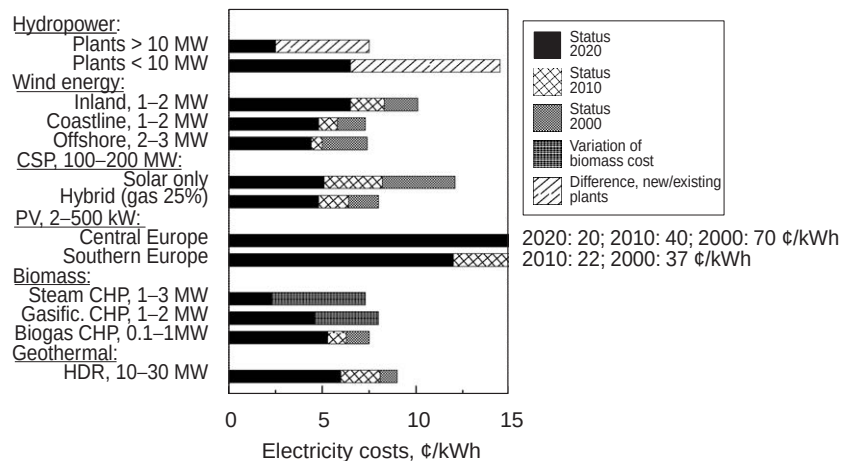


FIGURE 9 Survey of current electricity generation costs (total length of bars) of representative renewable energy source technologies and possible cost reduction until 2010 and 2020 (CSP, concentrating solar power; PV, photovoltaics; CHP, combined heat power; HDR, hot dry rock).

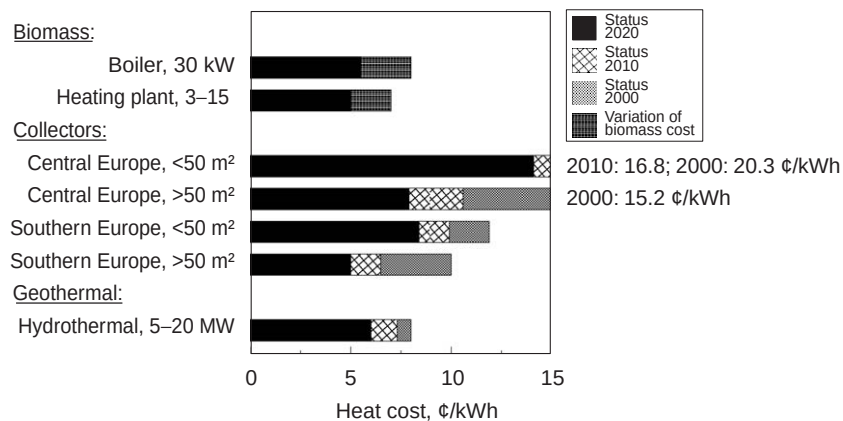


FIGURE 10 Survey of current heat production costs (total length of bars) of representative renewable energy source technologies and possible cost reduction until 2010 and 2020.

plants may generate power within the same cost range because of increasing fuel prices or, in the case of coal, because of costly measures for CO₂ separation. Under such conditions, the many RES technologies could be competitive economically. Photovoltaics, however, at 13€/kWh in Southern Europe and 20€/kWh in Central Europe, will still be too expensive to be used to a large extent for general electricity supply. The cost of biomass technologies will strongly depend on the price of the available biomass. Cost reductions by means of technological development or mass production will be small, because of stricter emission standards. If biofuels are to be used to a larger extent, the cost of electricity will not likely decline below a level of about 7–8€/kWh. Therefore, it has been proposed to use biofuels preferably for the production of transportation fuels,

although this will be more expensive than using gasoline or diesel in the foreseeable future. Because of relatively high taxes on fossil fuels in Western Europe, biofuels, with production costs of about 2.5–3€/MJ, are competitive, if they are not taxed. This approach has been pursued in Germany since 2001 and has resulted in a substantial increase in sales.

Costs of RES technologies for heat production can be assessed similarly. RESs can provide heat with stand-alone systems (e.g., wood-fired boilers, small solar-thermal collectors) and by means of district heating networks. District heating networks are of outstanding importance for RESs to achieve a breakthrough in the heat market. In most cases, only district heating needs support the use of RESs, e.g., geothermal heat plants, large solar collector fields for space heating, and biomass-fired cogeneration plants.

Costs of heat distribution in typical Western European settlements are in the range of 23¢/kWh. Medium-sized heating plants and large collector fields require lower specific investments compared to small boilers or small collectors. Heating costs for the end user of both configurations are therefore comparable (Fig. 10). RES heating costs show large differences. Water heated by typical solar collectors costs about 20¢/kWh in Central Europe; larger collector fields are able to reduce the costs to 15¢/kWh, but this is still four times the price of warm water from an oil- or gas-fired boiler (actual cost of 5–6.5¢/kWh). Only biomass heating is competitive today, if wood is available at a very low cost. Warm water solar collectors in Southern Europe are less expensive (10–12¢/kWh), making them competitive with electrical boilers. Future collector systems will be able to reduce heat costs considerably. Particularly, large collector fields, covering 100 to more than 1000 m², will allow future costs to be about 7¢/kWh. The direct use of geothermal energy will result in future costs of about 6¢/kWh. In view of the likelihood of rising prices of fuel oil and natural gas in the future, biomass plants, geothermal units, and solar collector systems will be able to supply end users with competitive thermal energy.

4.3 Environmental Characteristics of Selected RES Technologies

The environmental problems arising from the use of fossil and nuclear energy sources are well acknowledged today; these include global climate change, acidification of ecosystems, risks from nuclear accidents, problems with long-term storage of radioactive waste, and public health effects from air pollution. Because RES technologies largely rely on natural energy and material flow cycles, they offer significant promise in reducing the environmental impact of energy use. Although for most of the RESs the energy conversion process is practically emission free, there may be environmental impacts from the development of RES technologies. The following environmental characterizations of current RES technologies for electricity generation, and their comparison against a modern natural gas-fired combined cycle (CC) power plant, are thus based on a life cycle perspective, taking into account emissions from all of the upstream and downstream processes related to energy conversion (Fig. 11).

It is obvious that for most of the RES technologies, consumption of nonrenewable energy resources and

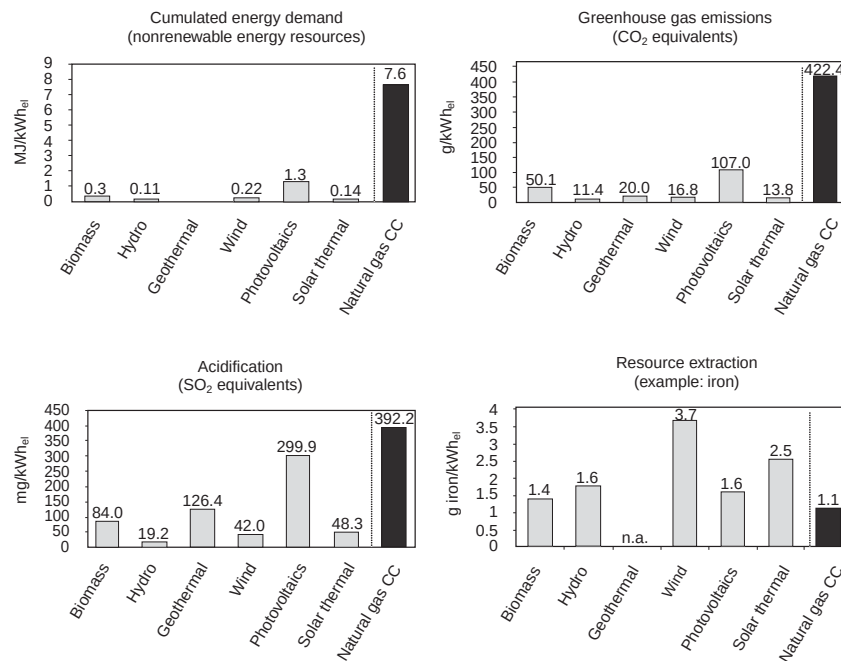


FIGURE 11 Environmental characteristics (cumulated energy demand, greenhouse gas emissions, acidification, and resource extraction) of renewable energy source technologies for electricity generation. Biomass (forest residuals), 20-MW steam turbine; hydro, 3-MW run-of-river plant; geothermal, 900-kW organic Rankine cycle cogeneration plant; wind, 1.5-MW onshore plant; photovoltaic, 3-kW p-type Si roof application (Central Europe); solar thermal, 80-MW parabolic trough (Southern Europe); natural gas CC (combined cycle), 58% efficiency.

greenhouse gas emissions are more than an order of magnitude lower than for the highly efficient gas-fired power plant; the difference is even considerably larger compared to less efficient coal-fired power plants. This fact underlines the important role of RESs in climate protection. Electricity generation from hydro-power, wind, and solar-thermal power plants ranks particularly high in this regard. For geothermal energy, the combined production of heat and electricity is required to achieve similar values. The harvesting and processing of biomass fuels requires substantial combustion of fossil fuels, thus biomass electricity generation ranks slightly worse than the processes not requiring combustion. Despite significant improvements in recent years, the manufacturing of photovoltaic cells still requires a quite high material and energy input, leading to relatively high life cycle emissions, which, in the case of SO_2 , are even in the same order of magnitude as emissions from gas-fired power plants. The example of iron ore indicates that for the RES technologies, the demand for nonenergy resources is higher, compared to the fossil option, a fact that is sometimes considered a critical drawback of renewable energy sources. However, the expected increase in recycling rates, which in most cases goes along with a further reduction of environmental impacts, compared to the extraction of primary raw materials, will eventually lead to a significant reduction of primary resource consumption.

The dominant environmental impacts of RES technologies are related to emissions from energy conversion in upstream processes, such as component manufacturing or transportation. The data thus primarily characterize the resource efficiency of the underlying economy rather than the performance of a specific energy conversion technology. Changes in the national energy mix, for instance, have a direct impact on the life cycle emissions of RES technologies. The evaluation of emerging technologies should thus be based on conditions that are representative for the time of their market entry, rather than associating them with the environmental load from technologies that they are expected to replace. Figure 12 illustrates the large potential for reduction of life cycle CO_2 emissions from a photovoltaic roof application due to key technological developments (use of solar-grade silicon, increase in recycling rates, and electricity generation with a high share of RESs).

Environmental impacts from energy generation include public health effects, yield losses of crops, and material damage, posing on society a significant economic burden that is not reflected in current energy prices (external costs). “Getting prices right” means imposing a surcharge on the energy price that accounts for environmental damages from power plant emissions. Although quantification of environmental impacts and the subsequent monetary valuation are extremely complex and a matter of

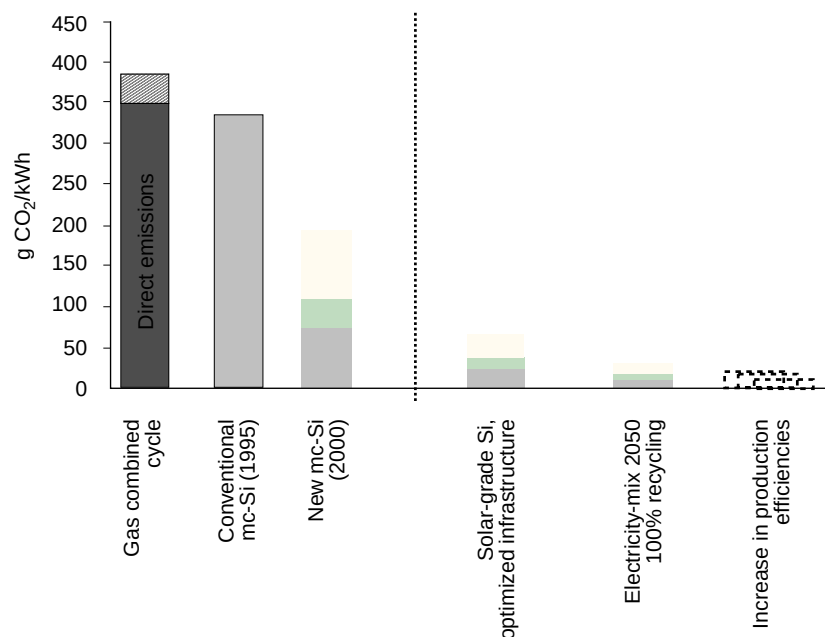


FIGURE 12 Life cycle emissions of CO_2 from a 3-kW p-type Si roof application as a function of technological development and change in the energy mix of an economy. mc-Si, Multicrystalline silicon.

many uncertainties, the European Commission, based on a thorough research effort, decided to allow member states to grant operating aid of up to 5¢/kWh, on the basis of external costs avoided, to new plants producing renewable energy. This measure for internalizing environmental damage costs underlines the environmental advantages of RESs and helps to further increase the competitiveness of RES technologies based on environmental considerations.

5. PROSPECTS OF RESs IN EUROPE

In Europe, several goals have been set to increase the use of renewable energy sources:

- In December 1997, the European Commission adopted the White Paper for a community strategy and action plan, “Energy for the Future: Renewable Sources of Energy”. The objective is to increase RESs to an amount equal to 12% of the EU gross inland energy consumption by 2010. In 2001, this action plan was supplemented by a directive of the European Parliament on the promotion of electricity from RESs. The target is to increase the share of RES electricity generation to 22% of total consumption in 2010. The directive contains specific targets for the individual share of RES electricity per EU member state.

- In 1999, the European Commission started a Campaign for Take-Off with the intention to start the implementation strategy set out in the White Paper with specific goals for the period 1999–2003. In 2001, an additional draft directive on biofuels was proposed. The aim was to increase the consumption of biofuels to 2% of the consumption of diesel and gasoline in 2005.

In fact, RES production and use have accelerated clearly since 1995. Between 1990 and 1995, the average growth rate of RES primary energy production in the EU-15 was 2.2%/year; in the following 5-year period, it rose to 3.6%/year and the current growth rate amounts to 4%/year. The average growth rate of the past decade was 2.9%/year. However, the targets defined in the White Paper demand an average growth rate of 7.3%/year until 2010, 2.5 times the value of the past decade (Fig. 13); a somewhat lower increase is necessary for electricity generation (Fig. 14).

A recent investigation, having monitored the progress of RES goals, indicates that the White Paper targets will not be reached with the present

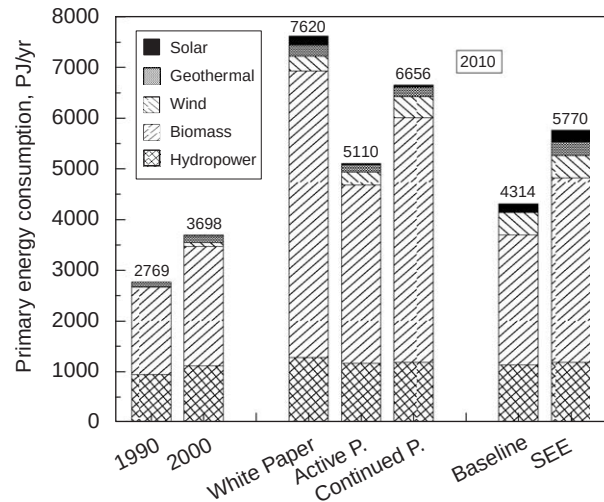


FIGURE 13 Contributions of renewable energy sources to primary energy consumption in the EU-15 in the past (1990, 2000) and to meet the goal for 2010 (from the EU Commission 1997 White Paper), and possible contributions in 2010 for different scenarios [active/continued P (active/continued policy); SEE, solar energy economy].

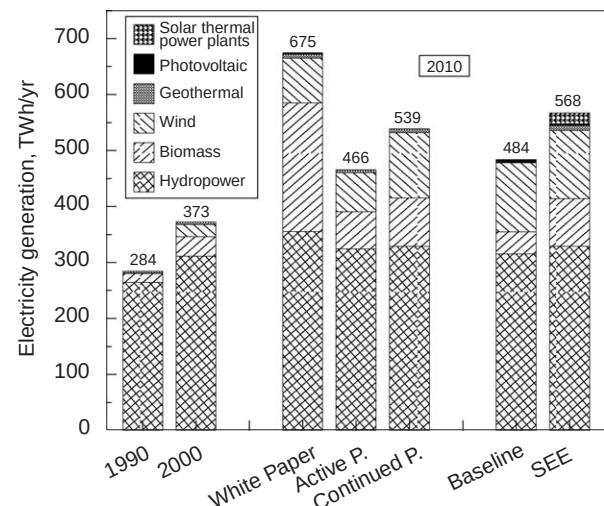


FIGURE 14 Contributions of renewable energy sources to electricity generation in the EU-15 in the past (1990, 2000) and to the goal for 2010 (from the EU Commission 1997 White Paper), and possible contributions in 2010 for different scenarios [active/continued P (active/continued policy); SEE, solar energy economy].

policy instruments and measures. The data on active policies (Figs. 13 and 14) for the year 2010 point out that only an estimated two-thirds of the target values will be achieved in time. All approved policy instruments that have passed in the parliaments of the different EU member states before September 1,

2001 have been considered. Assuming that currently existing incentives will be active in a foreseeable time (continued policy; Figs. 13 and 14), probably 80–85% of the target value will be achieved. Only the target value for wind energy will clearly be exceeded, with 60 GW of installed power in 2010 (White Paper goal: 40 GW); all other RESs will remain below the original objectives. The most severe discrepancies are expected with objectives for photovoltaics (only 20% achievement) and biomass electricity generation (35% achievement).

Substantial structural changes in an energy supply system always require long time periods. Therefore, it is of crucial importance to take into account long-term considerations concerning the transformation of the EU's energy supply structure in order to describe the prospects of RESs appropriately. The November 2000 Green Paper of the EU-Commission, "Towards a European Strategy for the Security of Energy Supply," is based on a projection of the European energy system up to 2030. Hence it can be used as a guideline for investigating the future role of RESs in Europe. The reference scenario baseline uses a business-as-usual development of the demographic and macroeconomic assumptions and is based on the EU energy policies currently in place. The gross domestic product of the EU will be 80% higher in 2030; population will rise only modestly, by 12 million people, until 2010 and will be stable thereafter, with 384 million people. Price increases for energy are assumed to be moderate until 2030, as exemplified by the crude oil price of \$30/bbl (U.S.) in 2030. Policy decisions on nuclear energy remain unchanged throughout the projection period. This results in a decrease of the contribution of nuclear energy to half of the present value until 2030. These basic conditions lead to an increase in primary energy demand of 11%, to a growing dependence on energy imports from the present 49% up to 70%, and (because of growing coal consumption) to an increase in CO₂ emissions of 22%, compared to the 1990 reference value of 3068 Mt/year. Apparently, these assumptions also include a weak growth of RESs. Until 2010, RES growth roughly corresponds to the data derived from "active policy" (Figs. 13 and 14). Considering the current dynamics of wind energy growth, only the electricity generation is somewhat higher (baseline, Fig. 14), but the growth of the primary energy contribution of biomass is estimated to be clearly below the value in active policy data (baseline, Fig. 13).

Against this background, it has become clear that the targets of the White Paper for the RES contribu-

tion in 2010 cannot be met in time if only the present energy policy instruments (described in Section III) are taken into consideration. Also, after 2010, the growth of RESs will not gain substantial momentum. Hence the European energy policy has to undertake further supportive activities in the next few years to accelerate the market penetration of RES technologies. From a current point of view, the solar energy economy (SEE) scenario (Figs. 13 and 14) gives a realistic assessment of the RES contributions that can be mobilized within the next 8 years. With the strategy "continued policy" as a starting point, it is assumed that the existing energy policy instruments will be strengthened and coordinated until 2005 in order to effect accelerated growth of RESs before 2010. In this way, almost all goals of the White Paper can be reached until 2010, except for the very high rate of biomass electricity generation assumed in the White Paper for 2010 (230 TWh/year compared to 35 TWh/year in 2000) and the somewhat lower growth rate of solar-thermal collectors. As a complement, electricity generation by concentrating solar-thermal power plants in Southern European countries is added in the SEE scenario; this has not been considered in any other scenario up to now. It is assumed that CSPs will generate 20 TWh/year electricity in 2010, corresponding to 5000 MW of installed power. This takes into consideration the legally mandated improved conditions in the Spanish electricity feed, which allows economically feasible operation of CSP plants by means of a guaranteed payment of 16¢/kWh. Several projects are in preparation now. In the SEE scenario, total electricity generation by RESs will equal 568 TWh/year in 2010, a share of 20% (Fig. 14). The contribution to gross primary energy demand increases to 5770 PJ/year, which is equivalent to 9.5% (Fig. 13).

The long-term development potential of RESs and their significance for a future energy supply system depend decisively on the status and the growth dynamics that have been reached by 2010. Comparing different long-term scenarios for the energy supply of the EU, this is quite obvious. It is noteworthy that the already discussed reference scenario "baseline" shows only a small growth of RESs over several decades (Figs. 15 and 16). The share of RESs for electricity generation obtains the 2010 target of the White Paper in 2030, and the growth in the heat and fuel market almost stagnates, so that the RES contribution to gross primary energy consumption rises only marginally. Extrapolating this scenario to 2050, RESs will fulfill at best a share

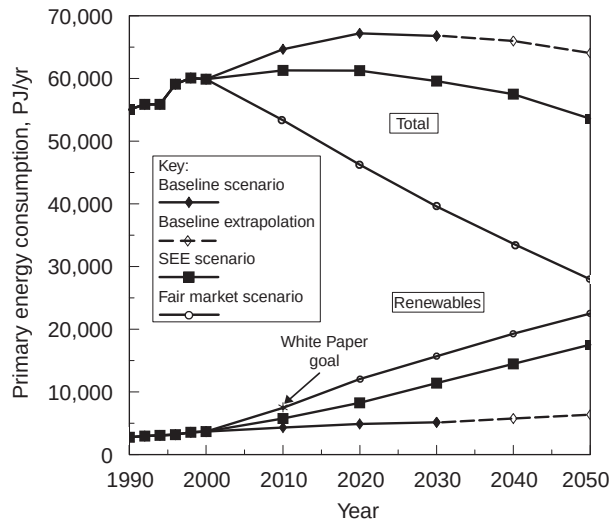


FIGURE 15 Future primary energy consumption in the EU-15 until 2050 and possible contributions of renewable energy sources for different scenarios. SEE, Solar energy economy.

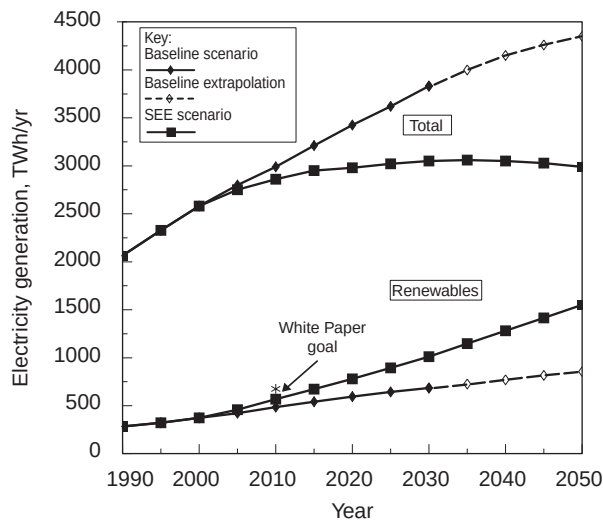


FIGURE 16 Future electricity generation in the EU-15 until 2050 and possible contributions of renewable energy sources for different scenarios. SEE, Solar energy economy.

of 10% of primary energy consumption (6400 PJ/year) and would remain a nearly negligible energy source. Such business-as-usual development would offer no chance for RESs to contribute significantly to climate change mitigation or prolongation of oil and gas resources.

A totally different perspective is offered by the fair market scenario (Fig. 15) that was developed by a consortium of five research teams from Germany, France, Belgium, and Denmark in 1997. The main

approach of this scenario is the inclusion of externalities of energy consumption in energy prices to determine market behavior. Once the magnitude of external costs has been fixed, e.g., by a political consensus on taxation, the market can act on prices as usual. Adding the known externalities to market prices of energy leads to considerably strong increases and results in price levels that are twice to three times as high as the energy prices used in the baseline scenario. In parallel, learning curves have been used to calculate the decreasing future cost of RES technologies. With these conditions, a strong market growth of RESs as well as of energy-saving technologies can be expected. The transformation of the energy system will be determined mainly by structural limits, e.g., the decommissioning of power plants and boilers for heating, the reconstruction of buildings, the construction of new production sites for RES technologies, or necessary changes of infrastructure for new fuels. These very favorable conditions for RES technologies result in average growth rates for all technologies of 4.5%/year during the next three decades, starting with about 7%/year in the next 10 years. Growth is assumed to continue until 2050, with slowly declining rates, resulting in a RES contribution of 22,400 PJ/year in 2050 to primary energy consumption (Fig. 15), which corresponds to six times the present value. Energy demand declines simultaneously because many energy-saving technologies are cost-effective at the assumed prices. The fair market scenario shows that, in 2050, the demand for energy services in the EU-15 can be met by half of the present energy production. Therefore, RESs are able to contribute 80% of the remaining primary energy consumption. As a result, CO₂ emissions are negligible compared to the present value, and the dependency on external energy resources is diminished to 15% of the actual imports of crude oil and natural gas. This scenario describes a development of the European energy system that could occur under very favorable political conditions. It would considerably reduce the deficits of the current energy system with respect to sustainability. In fact, this scenario is rather unlikely from the present point of view, but it shows, in principle, the unused potentials of RESs and rational approach of modern economies.

The third scenario, SEE, tries to balance the challenges of reaching sustainability in energy supply and the well-known slowness in removing the numerous barriers and restraints, which are opposed to accelerated structural changes in the European energy system. The SEE scenario outlines the chances

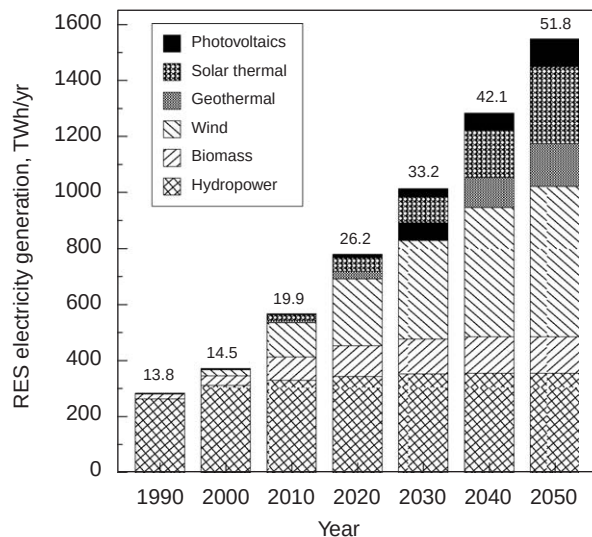


FIGURE 17 Growth of renewable energy source electricity generation in the solar energy economy scenario in the EU-15 until 2050, by individual sources. Numbers over bars indicate the percentage of total generation.

for a more independent, secure, and environmentally sound energy system in Europe with “careful optimism.” This includes an efficient and quick advancement and coordination of support mechanisms for RESs (and for energy conservation) in the EU, but, on the other hand, no excessive demands by society and economy. Starting from the interim values for 2010 shown in Figs. 13 and 14, the contribution of RESs grows continuously and meets in 2050 one-third of total primary energy consumption (17,500 PJ/year). Also, reinforcement of energy conservation measures has an impact, resulting in a total energy productivity in 2050 of twice the present value. In particular, it is assumed that the steady growth in electricity demand will be slowed down by adequate energy conservation efforts. In combination with increasing RES electricity generation, about 50% of electricity demand in 2050 can be met by RESs (Fig. 16). RES contributions to the heat and fuel market are growing somewhat more slowly.

To achieve an economically attractive growth of RES production and use, a well-balanced and well-timed mobilization of all RES technologies is of paramount importance. This mobilization depends on the technical potentials of the technologies, their actual costs, their cost reduction potentials, and their technological maturity. In Fig. 17, this is demonstrated with the RES electricity generation of the SEE scenario. Until 2010, cost-effective hydropower, biomass power plants, and wind energy converters

are the main contributors to RES growth. Hydropower and biomass will then reach their limits (biomass is increasingly used for fuel production). After 2010, the continuously growing use of wind energy is complemented by electricity from solar-thermal power plants and from geothermal energy. Relevant contributions from photovoltaics will not emerge until 2030 because of relatively high costs. The activities within the SEE scenario correspond widely to the minimum level of technically feasible changes. CO₂ emissions in 2050 are reduced to 2100 Mt/year, or 67% of the current value. This is only a first approach to an effective CO₂ abatement policy. Also, the dependency on imports of crude oil and natural gas, at 65% of the present value, is still too high. Altogether, the scenario represents the minimum requirements for European energy policy, if a relevant role of RESs in future energy supply is one of the substantial goals.

SEE ALSO THE FOLLOWING ARTICLES

Biomass for Renewable Energy and Fuels • Computer Modeling of Renewable Power Systems • European Union Energy Policy • Public Reaction to Renewable Energy Sources and Systems • Renewable Energy and the City • Renewable Energy in Southern Africa • Renewable Energy in the United States • Renewable Energy Policies and Barriers • Renewable Energy, Taxonomic Overview • Renewable Portfolio Standard

Further Reading

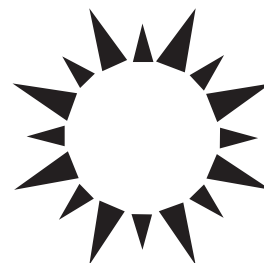
- Capros, P., *et al.* (2000). “Scenarios Related to the Security of Supply of the European Union.” Report for EU15. National Technical University of Athens, Greece.
- Commission of the European Communities. (1997). White Paper. “Energy for the Future: Renewable Sources of Energy.” EU-Commission COM (97) 599, final. European Commission, Brussels, Belgium.
- Commission of the European Communities. (2000). Green Paper. “Towards a European Strategy for the Security of Energy Supply.” EU-Commission COM (2000) 769. European Commission, Brussels, Belgium.
- ENER Forum. (2002). “Successfully Promoting Renewable Energy Sources in Europe.” The European Network for Energy Economics Research. Conference Proceedings, Budapest, Hungary, 6–7 June, 2002.
- European Commission. (1999). “ExternE—Externalities of Energy, Vol. 7: Methodology 1998 Update (EUR 19083). EU-DG XII: Science, Research and Development.” Office for Official Publications of the European Communities, Luxembourg.

- Eurostat (Statistical Office for the European Communities). (2001). "Renewable Energy Sources Statistics in the European Union. Eurostat, Theme 8: Environment and Energy, Edition 2001." Eurostat, Luxembourg.
- Hamelink, M., *et al.* (2002). Implementation of renewable energies in the European Union until 2010. "Study within the Project PETIR (Progress of Renewable Energy: Target Setting, Implementation and Realisation). Summary." Ecofys, Utrecht, Netherlands.
- LTI-Research Group (ed.). (1998). "Long-Term Integration of Renewable Energy Sources into the European Energy System." Physica-Verlag, Heidelberg and New York.
- Nitsch, J., and Staiss, F. (2003). Strategies for introducing renewable energies and the contribution of photovoltaics. In "Photovoltaics Guidebook for Decision Makers" (A. Bubenzer and J. Luther, Eds.). Springer, Berlin and New York.
- Nitsch, J., *et al.* (1999). "Climate Protection Using Renewable Energies" [in German; English summary]. Study commissioned for the Umweltbundesamt Berlin, UBA-Report 2/00. Erich Schmidt Verlag, Berlin.
- Orthen, S., and Urbschat, C. (2002). "Sunrise 2002—The European Markets for Solar Thermal Collectors and Photovoltaics" [in German]. German Energy Agency, Berlin.

Renewable Energy in Southern Africa

DIETER HOLM

Sustainable Energy Society of Southern Africa
La Montagne, South Africa



-
1. Impact of Renewable Energy in Southern Africa
 2. The Southern African Development Community
 3. Angola
 4. Botswana
 5. Democratic Republic of Congo
 6. Lesotho
 7. Malawi
 8. Mauritius
 9. Mozambique
 10. Namibia
 11. Seychelles
 12. South Africa
 13. Swaziland
 14. Tanzania
 15. Zambia
 16. Zimbabwe

Glossary

biomass Biological material (in a given area).

capacity The rating of a generator, transmission, or system.

desalination Salt removal from brackish or seawater.

geothermal Pertaining to the heat of the earth's subsurface.

photovoltaic (PV) A device converting sunlight into electricity.

resources Estimated total quantity such as minerals in the ground.

Renewable energies (REs) are those naturally occurring nondepleting energies that are derived directly and indirectly from the sun, such as biomass, geothermal, hydropower, ocean currents, solar radiation, tidal wave, and wind energy. These can be harnessed through renewable energy technologies

(RETs), such as photovoltaic, solar water heaters, wind and water turbines, and solar passive design. RETs deliver energy services such as mechanical work, heating, cooling, lighting, crop drying, telecommunication, transport, water pumping, desalination, and water treatment. REs are contrasted with finite nonreplenishing fossil fuels such as coal, petroleum, and uranium. REs and RETs are closely linked to new technologies such as fuel cells and the hydrogen economy as well as to energy efficiency. Africa relies heavily on traditional renewable energy in the form of fuelwood (70–80%) for pest control, water heating, space heating, food preparation, and beer brewing for the men. Beer brewing requires 25–50% of energy in villages. Women and children collect the fuelwood over ever-increasing distances, following the frontiers of desertification. The current and future impact of RE in southern Africa is a major building block toward sustainable development.

1. IMPACT OF RENEWABLE ENERGY IN SOUTHERN AFRICA

Solar cooking and solar water heaters (SWHs) can eliminate the need for collecting fuelwood and inhaling unhealthy smoke. This contributes to social sustainability because women and children who currently spend excessive time and energy collecting wood have time for education. Education and training contribute to improved work opportunities and a lower fertility rate, thereby advancing economic sustainability. The reduced population and renewable energy technologies (RETs) eliminate deforestation and desertification, thereby fostering environmental sustainability.

Solar passive buildings in urban and rural areas can totally obviate the need for artificial space

heating and cooling by kerosene, which often is a cause of death by fires or accidental swallowing. It also eliminates the use of expensive imported electricity and its environmental impact where this is generated by coal, nuclear, or hydropower schemes larger than 10 MW.

Modern RETs contribute to energy supply diversity, minimizing cost exposure to imported fossil fuels. Most African countries are importers of petroleum.

REs have a much lower cradle-to-grave life cycle impact on the environment than conventional fossil fuels. The environmental costs (social costs) of conventional energies are those costs caused by the generator of energy but paid for or carried by the local, regional, or global community, including 2 billion people who have no access to modern energies.

REs can be generated locally, feeding the electrical grid and reducing network losses. This is significant in southern Africa because power generation is centralized and power is transmitted over large distances.

Distributed RE generation also democratizes power, making it less likely that politicians can manipulate energy. Localized RETs enhance ownership of local communities that receive the local benefits and revenues. RETs are a new technology creating new export potential.

RETs create many more jobs than conventional energies. Job creation is a major instrument in poverty alleviation. The European Union plans to create 900,000 new jobs by 2020. Southern Africa may create approximately 1 million new work opportunities.

Shell International foresees that REs will provide more than 50% of the world's energy. By employing modern RETs, southern Africa's RE share may be 20% by 2020 and 50% by 2050.

2. THE SOUTHERN AFRICAN DEVELOPMENT COMMUNITY

This section discusses salient characteristic communalities under the headings Natural/Physical, Institutional/Political, Social, Economic, and RE Technological. The ensuing country-specific sections are structured in the same manner to simplify cross-referencing. Although every effort was made to collate recent and credible information, such as that supplied by Dr. P. Ashton (personal communication, 2003), it should be borne in mind that statistics for the Third World generally tend to be dated and unreliable.

2.1 Natural/Physical

The Southern African Development Community (SADC) is a group of 14 countries consisting of Angola, Botswana, the Democratic Republic of Congo (DRC), Lesotho, Malawi, Mauritius, Mozambique, Namibia, Seychelles, South Africa, Swaziland, Tanzania, Zambia, and Zimbabwe.

The country with the largest area is the war-torn DRC (2,345,410 km²), which also has the largest population (52,046 million). The SADC countries extend over sub-Saharan Africa from 5° north to 34° south latitude, from the equatorial tropics over the arid Horse Latitudes to the Mediterranean climate zone of the Cape of Good Hope. Although the Southern Hemisphere receives intensive sunshine during summer (perihelion), temperatures are tempered by altitude and cloud cover. 59% of the world's most intense winter sunshine (more than 6 kWh/m²/day) falls on southern Africa (for comparison, London receives 1 kWh/m²/day). The climate is characterized by steady hot and humid conditions in the Congo Basin and major temperature and humidity changes in the remaining inland zones, which are not tempered by the sea. Snow only occurs in the eastern and southern mountain ranges. The warm, low-altitude areas are malaria infested.

2.2 Institutional/Political

The forerunner of SADC was founded to destabilize South Africa during the Apartheid era. Today, SADC includes South Africa and has the aim of regional cooperation and economic development. South Africa's dominant technoeconomic power understandably makes for very unequal partners.

The New Partnership for African Development (NEPAD) proposed by South Africa's president Thabo Mbeki offers a promise that African rulers will in the future govern their countries according to rule of law in return for \$64 billion in foreign investments and donations. However, potential international long-term investors and donors have taken a wait-and-see position in view of the situation in Zimbabwe and Mbeki's efforts at reconciling the demands of the poor with those of development.

The NEPAD office identified six conventional electricity projects, two of which are in SADC: a 1300-MW hydropower scheme downstream from the Cahora Bassa dam in Mozambique and an interconnector between Malawi and Mozambique. Ironically, modern renewable energy projects are not

part of the NEPAD wish list, although Africa enjoys 96% of the world's best sunshine.

An African power pool has also been proposed by South Africa's state monopoly utility. The World Bank, the African Development Bank, the West African Development Bank, and others have shown interest in a proposed West African power pool, which would involve several major electrical power stations and interconnectors at a total cost of \$13 billion. However, international investors remain lukewarm because of the region's tradition of military coups.

The often quoted nominal 39,000 MW Inga hydropower, which actually could produce a maximum of 26,000 MW, has been criticized with regard to cost, time, and technical considerations, e.g., untested long-distance high-tension transmission technology. In addition, there are grave concerns about the risks in the conflict-ridden DRC.

The previously discussed schemes tend to have common traits:

1. With a view to attracting international finance with associated high transaction costs, the proposed schemes are very large by African standards. This means that African contractors are out of the game, except for performing menial tasks. At best, South African contractors would participate in worthwhile work, thereby enlarging the existing continental discrepancy.

2. All schemes are based on the dated concept of supply side-driven interventions typically promoted by large international finance, engineering, and Enron-type utility interests combined with local political positioning. The modern concept of need-driven integrated resource planning with its concomitant developmental benefits is conspicuously absent.

3. The schemes are of very large scale, tying up huge amounts of money and other resources that are far beyond Africa's ability to absorb the electricity in productive enterprises. A system of smaller, localized, modular, and incremental renewable energy programs would be much more appropriate due to Africa's immense distances and abundant renewable resources.

4. The schemes blithely ignore Africa's state of political instability, which renders centralized electricity generation systems with long-distance transmission lines extremely vulnerable, as demonstrated by the Cahora Bassa hydropower scheme.

5. It is a tacit assumption of all schemes that large-scale electrification leads to development. Numerous projects throughout the world have shown the

opposite to be true: Electrification follows development. Gunter Schramm of the World Bank states that "electricity by itself does not provide new income sources. As has been shown again and again, electricity follows, rather than leads economic development."

6. Although megaelectricity schemes are important to some rulers they often lead to the removal of national protectionism thereby opening local markets to overwhelming global competition. This global competition has often grown by a combination of national subsidies like the U.S. and EU agribusiness, and by national protection like the current U.S. steel industry. The developing world perceives the World Bank and International Monetary Fund as pontificating: "Do as we say, not as we did and do."

2.3 Social

As noted by Ashton, the total population within SADC is approximately 201 million. The DRC has the largest population (52 million). Of the SADC total population, 34% is urbanized, 71% literate, 43.5% malnourished, and 9% of adults are HIV/AIDS infected. A total of 15.7% have malaria, leading to 300,000 deaths each year. Approximately 500,000 have tuberculosis, causing approximately 100,000 deaths per year. Of the two-thirds majority living in rural areas, 51.1% have no access to improved water and 52.1% no sanitation, whereas in urban areas these percentages are 10.8% and 21.9%, respectively.

There are approximately 400 languages, not dialects. South Africa has 11 official languages. The number of religious denominations is estimated to be 10,000, with 6000 in South Africa.

The cultural patterns range from prehistoric hunter-gatherers, represented by the San, to iron-age herdsmen, burn-and-slash migrant peasants, sedentary subsistence and commercial farmers, industrialists, and contemporary postindustrial cosmopolites.

Not surprisingly, the Human Development Index (HDI) varies widely from approximately 0.3 to approximately 0.7. Likewise, the HDI ranking is between 120 and 170.

2.4 Economic

SADC's total per capita income was \$570 billion in 2000, yielding an annual gross domestic product (GDP) per capita of \$2833. This ranges from approximately \$7600 in Botswana to an unknown

figure in Zimbabwe, where the inflation rate is 400%. Two-thirds of the population lives below the poverty line of \$2 per day, ranging from 40% in the Kingdom of Swaziland to more than 80% in the Republic of Zimbabwe. The GINI coefficient (the gap between highest and lowest 10% income) is 0.58, with a range of 0.38–0.70. The sub-Saharan power sector received \$363 million during 1990–1998, representing a mere 0.8% of the total \$46.39 billion foreign direct investment flow to the entire developing world.

Most economic indicators have been declining. A limited number of scientific publications, patents, and inventions of SADC mostly stem from South Africa. The mineral resources of SADC are in stark contrast with the economic output. Gold, diamonds, copper, zinc, iron, gems, and rare metals are mined on a large scale.

Gas has been found along the western, southern, and eastern coasts as well as in inland Botswana. From Angola northward, oil is being extracted. Open cast coal mines are found in large tracts delivering the feedstock for more than half of Africa's electricity and to Sasol. In addition, cost-effective renewable energy resources by far exceed current and projected needs.

Despite the abundance of raw materials and sustainable energy resources, SADC has not learned to benefit the resources, remaining largely dependent on expensive imported value-added products and know-how.

2.5 RE Technological

There are 33 telephone lines, 39 cell phones, 39 TVs, 253 radios, and only 18 Internet users per 1000 people in SADC. The number of Internet users range from 0.1 (DRC) to 71 (South Africa) per 1000. The paved road length per person ranges from 3.09 km (Namibia) to 0.02 km (DRC), with an average of 0.9 km. Whereas Botswana has 0.54 km railway length per 1000 inhabitants, Lesotho has zero. The SADC average is 0.34 km. It must be borne in mind that after waterways, railways provide the most energy-efficient transport mode. Road transport is relatively inefficient at the higher altitudes of southern Africa, and petroleum products are often imported.

It appears that SADC countries, with the exception of South Africa, have invested relative little in conventional energy-intensive infrastructure such as centralized coal-fired or nuclear power stations with transmission networks. Furthermore, two-thirds of the population lives in widely dispersed rural areas

and has low energy consumption needs and very low income levels, which do not justify grid extension. Extending the national grid into these areas would place an unrealistic burden on both the national economy and the recipients who are supposed to be beneficiaries.

Given the existing low investment in old fossil-fueled infrastructure, the very high cost and risk of extending this into thinly populated, low-consumption rural areas, and the abundance of renewable sources such as sun, wind, biofuel, and microhydro power, it seems obvious to change to clean modern and environmentally friendly technologies. However, a number of donor-driven projects introducing 50 W_p photovoltaic solar home systems (SHSs) in remote rural areas have led to disappointing results. The reasons are not well publicized:

1. SADC governments tend to have a centralist approach and try to channel all initiatives through government departments. Even in the developed world, this is not necessarily the most efficient and transparent process.
2. Localized government interventions with subsidized SHSs destroy the private enterprise market and cause suspicion of corruption.
3. Inferior products and underdeveloped technologies were introduced because there was no willingness or capacity to control quality.
4. Installation and especially maintenance proved to be extremely difficult in remote rural areas, where telecommunication and transport infrastructure is problematic, if not totally absent.
5. Existing utilities viewed SHSs as a threat and prohibited SHS developments near their power lines or suddenly brought the grid to areas where SHSs had been established. They also created the expectation, supported by politicians, that universal grid-connected power is a reality in the very near future. This established a reluctance to invest in SHSs.
6. In hindsight, it appears that the introduction of a novel technology into an impoverished market under the most trying physical conditions was not a good idea.
7. As a result of insufficient or misleading information, SHS users had unrealistic performance expectations, which led to the stigmatization of SHSs as the poor man's power. The rich role models were not seen to be using SHSs. It seemed that an experiment was being conducted on the poor, who could least afford it.
8. Experience in Europe and Japan demonstrated that the major growth in RE technologies such as

SHSs is in grid-connected applications within the framework of a RE grid-feeder law. This law encourages reverse metering of RE-generated electricity fed into the grid at higher tariffs than the fossil fuel grid tariffs. The difference is shared by all grid users and is negligible. Batteries—normally the Achilles' heel of SHSs—are not required because the grid functions as the storage medium. This reduces capital and maintenance costs. Additionally, maintenance is cheaper due to easier access and communication. The well-to-do more readily introduce new technologies and demand service. All this contributes to a growing acceptance and desirability of photovoltaic (PV) systems as a status symbol of success and progress. So the mass market becomes established, leading to products and services that are acceptable and affordable to the less fortunate. After all, electronic watches, radios, and cell phones were not initially introduced into the low-cost end of the market. Learning from their own experience and from abroad, SADC countries can alleviate poverty and initiate sustainable development by initially introducing RE technologies using the successful and proven grid-feeder tariff structure. This will establish indigenous business, create jobs, and transform the market rapidly and sustainably. In this way, governments do not delay progress by interfering with implementation but, rather, create a conducive framework.

Most RETs are imported because of a lack of awareness on the part of decision makers and the public. R&D investment in RETs, energy efficiency, and government incentives have been neglected. The potential of clean development mechanisms has yet to be realized.

3. ANGOLA

3.1 Natural/Physical

Angola is rich in oil, gas, and mineral resources. It also has considerable solar potential ($5.4 \text{ kWh/m}^2/\text{day}$ in midwinter) in the southern region, except the coastal areas. The potential of small hydro is noteworthy. Wind and ocean energy resources remain to be assessed.

3.2 Institutional/Political

Governance institutions and infrastructure have suffered severely during the prolonged war, which, strangely, did not disrupt petroleum production. This

industry accounts for 54% of the GDP and 83.5% of the government tax revenue. Rural areas are not linked to the electricity grid, which has been sabotaged extensively. There is strong government control over tariffs.

3.3 Social

The life expectancy in Angola is the lowest of all SADC countries (due to war and HIV/AIDS). Literacy is also the lowest.

3.4 Economic

Foreign direct investment (FDI) has been twice as much as in Tanzania and far more than that for all other SADC countries. It has gone into the oil business; the per capita GDP is the fourth lowest. Both the African competitiveness ranking and HDI are the lowest. No concerted effort toward RE or energy efficiency has been undertaken.

3.5 Technological

There is no awareness of modern RETs, nor is there any local production. The high illiteracy rate is a barrier to employment. However, the need for PV, SWH, crop drying, and small hydropower is staggering.

4. BOTSWANA

4.1 Natural/Physical

The landlocked country has diamonds, copper, nickel, and coal of high rock and sulfur content, and the winter sunshine in excess of $6.5 \text{ kWh/m}^2/\text{day}$ places it in the core area of the world record sunshine. This enables all direct thermal applications, such as solar chimneys, solar ponds, solar concentrating, SWH, and PV.

4.2 Institutional/Political

A yearly National Development Plan has resulted in clearly defined targets and budgets. The Ministry of Mineral, Energy & Water Affairs (MMEWA) formulates, directs, and coordinates policy through the Energy Affairs Division, whereas the Botswana Power Cooperation, a parastatal under MMEWA, is a vertically integrated electricity utility operating the coal-fired power station at Murule that delivers 58%

of the national grid electricity. The remainder is imported.

Botswana has enjoyed long-term political stability, applies the death penalty, has a low crime rate, and has established investor-friendly policies. The Botswana Technology Center, the Rural Industries Innovation Center, and nongovernmental organizations are active in implementing RETs.

4.3 Social

Botswana has the highest regional urbanization rate of 46% and third highest life expectancy of 67 years, although HIV/AIDS is prominent. It is ranked third in competitiveness, with a comparatively low literacy rate of 67%.

4.4 Economic

The per capita GDP is the third highest and has been consistently so for a number of years. Its HDI is 57.2%, ranking lower than that of the Seychelles, Mauritius, South Africa, and Swaziland. Grid electricity is relatively expensive, which encourages the use of capital-intensive RETs.

4.5 Technological

Botswana is the region's leader in the application of SWHs and has its own manufacturing industry. Problems of scaling and high salinity experienced with direct thermosiphon systems led to the use of indirect systems. These are installed mainly in government and institutional buildings. However, the market is booming, with 8000–10,000 m² installations per year (\$2.5 million/year) and an installed capacity of 50,000 m². This amounts to 0.053 m² per capita, which is low by world standards.

PV installations are in use in the telecommunications sector. Government initiatives drive the PV market since grid extension is the costlier option for smaller structures such as housing for teachers and the defense force, prisons, public buildings, and government buildings in rural towns. Rural energization by SHSs is gaining momentum. The current market is estimated to be 13.4 kWh/year (\$70,000/year), with an approximate growth rate of 12% per year.

There is some local manufacture of balance of system components. The Botswana Technology Center is active in energy efficient/solar passive buildings. This institution has also performed a wind mapping project.

5. DEMOCRATIC REPUBLIC OF CONGO

5.1 Natural/Physical

This tropical rain forest straddles the equator and is the most northerly SADC country. It is well endowed with mineral resources, oil, and gas, and the Congo River has considerable potential for hydropower at Inga (26,000 MW).

5.2 Institutional/Political

Unfortunately, since liberation the DRC has experienced conflicting foreign and local interests with regard to its oil and mineral resources and seemingly unending wars. This has destabilized the area, with ripple effects on the whole of Africa. Governance is in a survival mode. Statistics are not reliable.

5.3 Social

The DRC has the largest population in the SADC, with a life expectancy of 51 years and HDI of 43.1%.

5.4 Economic

GDP was \$230/capita in 1998, which is the second lowest in the region; there appear to be no recent statistics. The FDI in 1995 was estimated to be \$2 million. There is neither a global nor an Africa competitiveness rating.

5.5 Technological

The DRC does not publish any official documentation on RE. The often mentioned hydropower potential of the Congo River seems fraught with political uncertainty. In addition, the major consumer in the near term would be South Africa, which would require more than 3000 km of new power lines, in part over land-mined areas.

6. LESOTHO

6.1 Natural/Physical

The mountain kingdom of Lesotho is an enclave within South Africa and Mozambique. Owing to its altitude (3096 m), it is much colder than its latitude (30°S) would indicate. Snow is not exceptional. The existing Oxbow hydro scheme exports to South

Africa (with a planned additional capacity of 200 MW_p). The inhabitants' use of firewood has practically denuded the land of trees, although reforestation conferences are popular. There is ample microhydro potential. Daily winter sunshine is 5.5 kWh/m²/day.

6.2 Institutional/Political

The Lesotho Energy Master Plan of 1988 was revised in 1999, with the Lesotho Electricity Master Plan Strategy of 1996 as a subcomponent, updated in 1998. Owing to the sparse population and the mountainous nature, only approximately 3% of the population has access to grid electricity, which is planned to increase to 5% by 2003. The Lesotho Electricity Corporation is being privatized.

6.3 Social

Of the total population, 17% is urbanized, but there is a relatively high literacy rate of 78%. The life expectancy is 59.4 years, whereas the HDI is 53.5%.

6.4 Economic

The GDP per capita is \$920, and the Africa competitiveness ranking is 13.

6.5 RE Technological

As a result of the low grid electrification, dispersed population, high solar radiation, and deforestation, Lesotho is a good country for solar applications. This is limited by the low disposable income of the population and the lack of trained installers.

The high altitude causes frost conditions, necessitating special precautions with SWHs. Four small hydro systems exist (3268 MW), and there is potential for more. Approximately 60 kW_p of PV has been installed, with limited growth prospects.

7. MALAWI

7.1 Natural/Physical

The country on the shores of Lake Malawi (706 m altitude) enjoys more than 5 kWh/m²/day of sunshine during the winter. The vegetation varies between savannah and tropical rain forest.

7.2 Institutional/Political

The Malawi legal system is based on English and local traditions. The energy policy stems from the Vision 2020 development plan, and the Electricity Council fulfills a regulatory function. A National Sustainable Renewable Energy Program was in place before the 2002 World Summit on Sustainable Development.

7.3 Social

Malaria, tuberculosis, HIV/AIDS, and continued malnutrition are the most pressing health problems. The literacy rate is 40% and secondary school enrollment 5%. The HDI is 40%, with a life expectancy of 52 years.

7.4 Economic

In African competitiveness, Malawi ranks 19th, with a GNP per capita of \$200. Foreign investment is approximately \$1.8 billion.

7.5 RE Technological

A total of 97.6% of Malawi's 214-MW grid electricity is hydro generated, and the government has borrowed \$8 million from the Development Bank of Southern Africa for extended hydroelectric power. Only 3% of rural households have access to grid electricity. Wood fuel continues to dominate the fuel economy, where households consume 70% of noncommercial wood fuel. The United Nations Educational, Scientific, and Cultural Organization (UNESCO) has assisted with a solar energy promotion program.

The installed capacity of PV is approximately 50 kW_p, increasing at approximately 4 kW_p annually. The potential is considerable. The SWH market is not very active. Approximately 3000 m² has been installed, amounting to a negligible market penetration.

8. MAURITIUS

8.1 Natural/Physical

This 2040 km² island is situated in the southerly warm Indian Ocean Current at approximately 20°S latitude, receiving 4.75 kWh/m²/day of sunshine in winter. The climate is tropical during summer (22–29° C) and subtropical in winter (18–23° C). This is

where the ebony forests and the Dodo became extinct. Rainfall varies from 1150 mm in the west to 2300 mm on the 580-m altitude plateau. Cyclones are not unusual.

8.2 Institutional/Political

The legal system is based on French and English law. Creole is widely spoken, with English as official language. In the absence of natural fossil resources, energy legislation is guided by the development of local RE resources and the involvement of the private sector, which utilizes hydropower and bagasse (sugarcane residues). However, the bulk is imported fossils. The energy demand is growing at 5% per year.

8.3 Social

The population of 1.1 million has a relatively high literacy rate of 81%, a life expectancy of 70 years, and the highest SADC urbanization of 43%. Consequently, the HDI is 77.2%, which is by far the highest in southern Africa.

8.4 Economic

Mauritius ranks second out of 24 in African competitiveness, and it is one of the few with a global competitiveness rating on record (32nd out of 75). The GNP per capita was \$ 3700 in 1998. Mauritius has high credit worthiness and was ranked first in Africa in 1996. Manufacturing, tourism, and financial/professional services lead the economy.

8.5 RE Technological

Diesel engines generate 61.7% (176 MW) of the electricity. A seasonal contribution of 54.2 MW comes from hydro plants and 90.5 MW from independent power producers, mainly bagasse plants. The entire population has electrical grid access. Mauritius is the SADC pioneer of grid-connected RE, including PV. Wind speeds of 8 m/s make wind turbines an attractive option. A total of 20,000 SWHs have been installed, representing a market penetration of 8% in the household market. The growth is stable at 2000 m² (\$1 million) per year.

More growth in grid-connected and stand-alone (on-island) PV is envisaged. Mauritius demonstrates that electrification follows development and that cheap energy is not a competitive advantage.

9. MOZAMBIQUE

9.1 Natural/Physical

Mozambique is a mostly low-lying country on the warm Agulhas Current of the West Indian Ocean, extending from 11 to 27°S latitude in a tropical to subtropical climate. Winter insolation is 5 kWh/m²/day (approximately twice the sunshine of Bilbao). It features an area of 799,380 km² and a coastline of 2515 km with two major African rivers—the Zambezi and Limpopo—traversing. The huge Cahora Bassa hydroelectric dam, which exports electricity, is on the Zambezi River. The Pande field northeast of Maputo will soon export natural gas to the petrochemical plants of Sasol in South Africa. There are also extensive coalfields at Metangula, Itule, Tete Msanghansuvu, and Chiomo.

9.2 Institutional/Political

After decolonization in 1975, a prolonged civil war began that ended in the 1999 general elections, leading to economic development from a low base. Mozambique ratified the SADC trade protocol and has the lowest import tariffs, along with Malawi. An energy-intensive aluminum smelter is under construction. The RE resources have not been assessed.

9.3 Social

Of the approximately 16.8 million people, 23% are urbanized. The life expectancy is 46 years, and literacy is 28%—the lowest in SADC. Portuguese is the official language. The HDI recently increased from 28.1 to 32.2%, whereas the African competitiveness rating is 18 of 24.

9.4 Economic

The joint Maputo corridor project linking South Africa's infrastructure to upgraded harbor facilities will boost the economy. Despite climate disasters, the Mozambique economy has grown remarkably. Foreign reserves increased from \$195.25 million in 1995 to \$652 million in 1999. FDI was \$4.07 billion in 1999.

Mozambique will own the Cahora Bassa scheme by 2014, producing 4000 MW upon completion of phase II.

9.5 RE Technological

Despite favorable solar conditions and only 7% grid connection, PV penetration is only approximately 175 kW_p, including 150 clinics. A study on PV was completed by the World Bank, but the document was not accessible. SWH is mostly limited to holiday resorts. Wind and small hydro IPP potential is estimated to be considerable, but it is unexplored.

Poverty and low energy demand resulting from civil war are barriers to short-term RE development. The joining of the Trans-Boundary Game Reserve with the existing South African Kruger National Park should offer RE opportunities.

10. NAMIBIA

10.1 Natural/Physical

Namibia, which has an area of 842,269 km², is bordered by the Atlantic Ocean's cold Benguella Current and has a mostly arid climate with semiarid and subtropical regions. Winter sunshine reaches 6.8 kWh/m²/day, the world's highest. The coastal strip is often misty during the night. Extreme daily temperature swings are typical. Namibia is very rich in diamonds (Sperrgebiet), other precious and semi-precious stones, ores, and uranium (Rössing). The perennial Orange and Cunene Rivers are the north and south boundaries. The Kudu gas field is situated off the coast between Oranjemund and Lüderitz.

10.2 Institutional/Political

The former German colony was taken by South African troops during World War II and became independent in 1990. The 1998 "White Paper on Energy Policy" refers to RE as an integral and equal component, with special reference to rural water supply and electrification. The restructured utility, Nampower, retains its generation and transmission asset, whereas IPPs provide new generation capacity. Similar to the proposed South African scheme, regional electricity distributors will be responsible for distribution. Namibia imports energy.

10.3 Social

The population of approximately 1.7 million amounts to only 1.7 persons/km², the lowest density of SADC. Approximately 27% of the population is urbanized, the literacy rate is 76%, life expectancy is 61 years, and there is a low HDI.

10.4 Economic

Namibia has a GDP/capita of \$1370 and is ranked fourth in the African competitiveness rating—the third highest in SADC. Its foreign reserves increased from \$221 million in 1995 to \$305 million in 1999. The main GDP shares are government services (22.4%) followed by mining (16.4%) and manufacturing (11.8%), indicating an overweight government.

10.5 RE Technological

Namibia has ideal climatic conditions for RE. Domestic SWH installations are in use in government buildings. However, the 24,000 m² installed capacity yields a very small market penetration of 0.001 m²/person and there remains great potential.

PV systems are applied in telecommunication, water pumping, and SHSs. Grid-feeder applications are not known. The total installed capacity is approximately 1 MW_p, with 446 kW_p in the household sector, which is expected to grow at approximately 15 kW_p/year (\$100,000/year).

Wind resources of 5–8 m/s along the coast led to advanced feasibility studies for a 5- to 20-MW grid-connected wind park at Lüderitz. This was turned down by Nampower. Traditional windmills for pumping water are gradually being replaced by PV.

11. SEYCHELLES

11.1 Natural/Physical

This archipelago of 115 islands totaling 455.3 km² is located 4° south of the equator in the convergence area of the warm monsoon drift and the warm southern trade wind current. The southeast trade winds blow from May to September. During the rest of the year, the northwest monsoon brings tropical rainfall, higher humidity, and temperatures up to 32°C. The area lies outside the cyclone belt.

The world heritage site Vallée de Mai is located on the island of Praslin and is the home of the world-renowned coco-de-mer and the endangered black parrot.

The annual rainfall of 2238 mm is high, but 98% is lost because of the steep topography and evapotranspiration.

11.2 Institutional/Political

Seychelles law is a hybrid of civil and English common law. An energy policy developed by the Ministry of Industry and International Business: Energy Affairs Division guides development, whereas the Public Utilities Corporation: Electricity Division is responsible for electricity generation, transmission, and distribution according to the traditional vertically integrated model.

11.3 Social

Most Seychellois speak Créole, French, and English, reflecting a colonial past, including Arabs and Portuguese taking up victuals and even a period of French and English pirates from 1685 to 1730. At 87.9% (1994), the Seychellois have the highest literacy within SADC and share the highest life expectancy of 70 years with Mauritius. The HDI is 81%, and there is no competitiveness rating.

11.4 Economic

Transport, distribution, and communications (28.3%) dominate the GDP, followed by manufacturing and handicraft (14.5%), government services (13.5%), and the growing finance and business services (11.8%). Fisheries and tourism reflect the traditional income base, which is changing from centralist control to greater private sector participation. The GDP per capita is \$ 7384 (1999)—the highest in the SADC region and 40 times higher than that of Malawi.

11.5 RE Technological

A total of 95% of the population has access to expensive grid electricity generated by heavy fuel oil instead of diesel. A 50-MW (\$4.8 million) power plant was built at Roche Caiman, approximately 2 km from Victoria.

Seychelles Petroleum provides oil and liquid petroleum gas (LPG). There is a government campaign to phase out kerosene for household cooking in favor of LPG. The search for hydrocarbon in the Exclusive Economic Zone continues under the Seychelles National Oil Company. High electricity costs led to the installation of more than 2400 m² SWHs yielding a market penetration of 0.03 m²/capita and growing by 400 m² (\$120,000) annually.

Steady wind regimes and high fossil fuel imports should lead to the use of offshore wind turbines.

Recent international experience demonstrated that grid-connected turbines feed 50% of the capacity into the grid. This translates into a wind energy potential of 25 MW.

PV SHSs are used on the islands. Biogas and biofuel are not being exploited because of a lack of awareness.

12. SOUTH AFRICA

12.1 Natural/Physical

This 1,221,000-km² country occupies the southern tip of the African continent between 22 and 34°S latitude. Its 2968-km coastline borders the Indian and Atlantic Oceans. Prince Edward and Marion Islands 1920 km south of Cape Town are also part of the territory. A mountain range along the southeastern coast limits precipitation in the interior, making the country semiarid with 464 mm of annual rainfall compared to the global average of 857 mm. Winter sunshine varies from 4.5 kWh/m²/day along the south and eastern coast to 6.8 kWh/m²/day in the northwestern inland areas. Good wind potential exists along the coasts and mountain areas.

12.2 Institutional/Political

The “Energy White Paper” was published by the Department of Minerals and Energy in 1999, and the “Draft White Paper on the Promotion of Renewable Energies Part 1” (begun in 1998 and published in Dec. 2003) while “Energy Efficiency Part 2” (begun in 1998) is expected to be published in the near future. The 1999 “White Paper” mandates breaking up of the vertically integrated electricity monopoly Eskom, which achieved the electrification of more than 2 million households since 1992. This restructuring of Eskom into the envisaged regional electricity distributors is imminent. An independent national energy regulator oversees energy licensing and tariffs, including electricity. A large rural off-grid energization drive using private sector monopolies was launched in 1999, and implementation is imminent. South Africa acceded to the Kyoto Protocol as a developing country.

12.3 Social

The population of 43 million is 30% urbanized, 82.2% (1991) literate, had a life expectancy of 66 years in 1991 that has been reduced to 47.7 years,

and has a HDI of 6.5%. Afrikaans, English, and nine official native languages are recognized as official languages. In addition, Portuguese and other European, African, and Asian languages are spoken.

12.4 Economic

Strict monetary policy is being applied in order to reduce the inflation rate. The exchange rate of the local currency (Rand) against the dollar decreased by 39.8% from 1995 to 2002, whereas the GDP per head increased by 6.4% at constant 1995 prices. Weakening of the U.S. dollar has resulted in an improved rate of 7.20 Rand per dollar. The international credit rating is 40, the global competitiveness rating is 34 out of 75, and the Africa competitiveness rating is 7 out of 24.

South Africa's energy intensity is 15%, which is nearly double the global 8%. Eskom is the world's fifth largest electricity utility (39,870 MW installed capacity), supplying 95% of the country's electricity, which is 95% coal generated.

South Africa is the world's third largest per capita CO₂ emitter, after the United States and Canada, contributing 1.7% to global emissions but only 0.4% to the global GDP.

12.5 RE Technological

RE energy has been neglected with regard to R&D and implementation. A national wind turbine demonstration site is being developed at Darling in the Cape and another test site has been started at Klipheuwel by Eskom.

A 200-MW agricultural upwind chimney delivering 24-h load-following electricity as well as agricultural produce is being considered in the Northern Cape Province. A feasibility study for a 100-MW solar thermal plant is under way by Eskom, which also erected an experimental dish sterling machine of 25 kW. That was inaugurated during the World Summit on Sustainable Development at the Development Bank of Southern Africa. The CSIR in Pretoria is investigating wave energy generators.

Currently, only the on-shore wind potential of the approximately 3000-km coastline is being assessed. If the wind energy contribution is capped at 50% of the present capacity, it amounts to approximately 20 GW, increasing to 25 GW in 2012.

The World Bank 10-year Global Environment and Local Development project is summarized in Table I.

The project includes the energy-efficient (solar-passive) design and construction of 300,000 low-

TABLE I

World Bank Renewable Energy Initiative in South Africa, 2003

	Installed/saved (MW)	Annual generation (GWh)	Estimated capital cost (\$M)
Commercial solar water heating	NA	690	510
Biomass cogeneration	220	1320	172
Paper and pulp waste	195	880	174
Municipal solid waste/landfill gas	125	825	145
Wind/other	60	247	56
Total	600	3962	1056

income houses. Currently, there are 184,000 m² SWHs, with a potential for 43 million m² by 2020, including domestic and commercial glazed and unglazed SWHs. There are no subsidies.

The announcement of the rural government-subsidized SHS program in 1999 comprising 250,000 units of 50 W to be built during a 5-year period created great expectations, but in early 2003 the Department of Minerals and Energy still had implementation problems. The total national unsubsidized PV capacity in 2003 was more than 8 MWh/year_p.

The bioenergy potential is given in Table II. The hydropower potential is 8,360 MW or 73,230 GWh/year, whereas the installed capacity was 68 MW in 2000. The potential wave energy is 56,800 MW. There is no installed capacity. The 2020 potential savings in energy-efficient buildings is 36,000 GWh/year.

Job creation is a major motivation for renewable energy application. Biodiesel can create more than 250,000 new jobs, energy-efficient buildings 250,000, PV 125,000, one solar chimney 43,000, solar water heating more than 15,000, and solar cookers 10,000. This totals more than 693,000 permanent work opportunities in a country with an estimated unemployment rate higher than 40%.

13. SWAZILAND

13.1 Natural/Physical

The Kingdom of Swaziland covers an area of 17,364 km² and is bounded by South Africa and Mozambique. It lies at the 26°S latitude on the

TABLE II
Bioenergy Potential^a

Bioenergy	2003 potential (GWh/a)
Bagasse	1400
Sawmills	7639
Pulp mills	4528
Agricultural wastes, manure, litter, grass	5615
Biodiesel	3,500,000
Bioethanol	1200
Total	3,520,382

^aFrom Sustainable Energy Society of Southern Africa (2002).

eastern side of the escarpment ranging from 98 to 1800 m in altitude. This causes orographic summer rains. The average annual rainfall is between 1016 and 2286 mm, which is high by southern Africa standards. Winter sunshine is more than 5.0 kWh/m²/day, with temperatures between 5 and 19°C, whereas summer temperatures vary between 15 and 25°C.

13.2 Institutional/Political

Swaziland's legal system is based on traditional Swazi national courts and constitutional courts derived from Roman-Dutch law. A national energy policy formulation project was completed by the energy section of the Ministry of Natural Resources and Energy assisted by Danish Cooperation for the Environment and Development.

The Swaziland Electricity Board (SEB) is a government-owned utility and provides 70% of the final commercial electricity, of which 30% is generated by SEB's hydro and diesel plants. The remainder is imported from neighboring South Africa's Eskom. Sugar and pulp cogeneration plants provide the 30% of the remaining national electricity supply not delivered by SEB.

A rural electrification policy envisages "that the access to energy is made available for all by 2010" and that "access to electricity is made available for all by 2020." This includes off-grid electrification. The primary focus is on schools, health care facilities, and other essential public facilities.

13.3 Social

English is the official language, with siSwazi spoken by the majority of the 1.02 million (1999) people.

The literacy rate is 78%, urbanization 34%, HDI 57.7%, and life expectancy 57 years.

13.4 Economic

The GDP per capita is \$1268 (1999), with the major contributors being manufacturing (36.9%); government services (15.3%); wholesale, retail, hotels, and restaurants (10%); and agriculture (9.8%). More than 80% of commodities, including petroleum products, come from South Africa. The African competitiveness ranking is 10 out of 24.

13.5 RE Technological

Renewable energy from bagasse and pulp cogeneration comprises a significant 30% of delivered energy. SWH is underdeveloped. PV systems are applied in an estimated total capacity of 50 kW_p, with approximately 400 domestic lighting systems (16 kW_p) and a growth of approximately 12 kW_p per year. With a 10% market penetration of stand-alone systems, the potential market is 600 kW_p. Two small hydro systems are in operation.

14. TANZANIA

14.1 Natural/Physical

This country is a merger of the former Tanganyika and Zanzibar, which comprises the Unguja and Pemba Islands in the Indian Ocean. In the west, it extends to Lakes Nyasa and Tanganyika and in the north to Lake Victoria. The world's largest crater, the Ngorongoro, rises to 3647 m.

The land lies between 1 and 11°S of the Equator and covers an area of 945,200 km². The floor of Lake Tanganyika is part of the Great Rift Valley and is Africa's lowest point at 358 m below sea level.

There are four main climatic zones: the hot and humid coastal plains; the hot and arid central plateau; the high, moist lake regions; and the temperate highlands. The northeast monsoon brings the hottest temperatures from December to March. Shifting south in March to May, it carries heavy intermittent rains. The southwest monsoon brings cooler, dry weather from June to September. The short wet season is interspersed in November and December. Annual rainfall is 1250 mm on the southeast slopes of the volcanoes and less than 600 mm in the central plateau.

14.2 Institutional/Political

Tanzania gained independence from British rule in 1961. The United Nations Development Program helped to develop the “National Framework for Renewable Energy Development” as part of the National Program for Renewable Energy. Tanzania is one of the few SADC countries with a national electricity regulator and an IPP structure.

14.3 Social

The estimated population is 32 million (1999), of which 30% is urbanized and 84% literate. The life expectancy is 50 years, with a HDI of 44%. KiSwahili is the official language. English is the business language mostly spoken by the 3% of urban Asians. Religions are Muslim, Christian, and traditional in approximately equal percentages. HIV/AIDS is a concern.

14.4 Economic

Tanzania has an old tradition in trading predating Arab slave, gold, ivory, and copper trading, which was eventually taken over by the British. Currently, agriculture accounts for 57% of the GDP, grown on arable land limited to 4% of the total area. The GDP per capita is \$242 (1999), and the Africa competitiveness rank is 7 out of 24.

14.5 RE Technological

There are gas fields (33 billion m³) at Songo-Songo and 1200 million tons of coal at Ufipa, Galula, Songwe, Ruhuhu, Songea, and Mkusi. Informal fuelwood provides 90% of all energy needs, petroleum provides 9%, and hydroelectricity and coal provide the balance. The SWH market is underdeveloped.

Approximately 500 kW_p of PV has been installed in clinics and schools in rural and urban areas. Approximately 60 small hydro plants (30 MW_p) have been installed, with an additional 50 identified. Approximately 1000 biogas plants have been installed, representing the most ambitious program in SADC. The follow-up is uncertain.

15. ZAMBIA

15.1 Natural/Physical

Zambia is landlocked in a tropical belt on a plateau at 1300 m covering 752,612 km². Summer temperatures range from 27 to 32°C, whereas the winter

range is 16–27°C. Solar radiation in winter is more than 5.0 kWh/m²/day, rendering it suitable for solar applications. The mighty Zambezi and Lake Kariba form the southern boundary, where it can be very hot during October. Annual rainfall during the warm, humid season from November to April is 508–1270 mm.

15.2 Institutional/Political

Zambia's borders are the result of Cecil Rhodes imperialism, which fell apart in 1964. In 1972, it became a one-party state until general elections were held in late 1991. The 1994 energy policy enjoys political support. There is a national electricity regulator and a reformed electricity system. A grid and non-grid electrification program is in place. Zambia has by far the best energy-efficiency organization in the SADC.

15.3 Social

Zambia comprises approximately 10 million people (1999) of the most diverse cultures who speak more than 70 dialects and have diverse lifestyles, including fishing, cultivating, and ranching. Christians comprise 80% of the population, and the rest believe in Traditional, Islam, Hinduism, and Buddhism. The literacy rate is 79%, and 38% of the population is urbanized, representing the highest percentage in the continental SADC. The life expectancy is a 46 years, and the HDI is 43.3%.

15.4 Economic

Zambia is the world's fourth largest copper producer and owns 6% of the known global reserves. Copper accounted for 50% of the country's export earnings in 1998. Zambia also produces more than 20% of the world's emeralds. Although the country received debt relief up to 67%, the debt burden remained unsustainable. The GDP per capita was \$302 in 1999, and the African competitiveness ranking is 12 out of 24.

15.5 RE Technological

Zambia produces 99% of its electricity by cheap hydropower used mainly in the mines, whereas only 20% of the population has access to grid electricity. Proven coal reserves are 30 million tons. Petroleum is wholly imported, draining approximately 10% of foreign exchange.

SWH has a minor role. Small-scale, stand-alone PV systems are in use with an estimated capacity of 1 MW. More than 20 small hydro systems have been installed and their number is increasing. IPPs offer good opportunities.

16. ZIMBABWE

16.1 Natural/Physical

Zimbabwe (390,575 km²) shares Lake Kariba with Zambia and is situated between the Limpopo and Zambezi Rivers, wholly within the tropics between 15 and 22°S at a high altitude of 800–1500 m above sea level. The average rainfall is between 196 mm in summer and 10 mm in winter. Most of Zimbabwe receives more than 5 kWh/m²/day in winter, representing a very good climate for solar applications.

16.2 Institutional/Political

Comrade President Mugabe is a prominent proponent of the World Solar Summit initiated by UNESCO. A major Global Environmental Fund PV program was launched in Zimbabwe. A rural electrification initiative is cross-subsidized by a levy on all electricity sales. Electricity is imported on a large scale from neighboring countries.

16.3 Social

Zimbabwe estimated its literacy rate to be 80% and its urbanization to be 31%. Life expectancy is 50 years, whereas the HDI is 55.1%. There is widespread starvation.

16.4 Economic

The GDP per capita is \$505 (1999), whereas the Africa competitiveness rating is 23 out of 24, and it was ranked 75 out of 75 according to the Global Competitiveness Report (2001/2002). The 2003 inflation rate was 400%.

16.5 RE Technological

Most of the 0.75-MW_p PV installations were donor funded, an activity that destroyed private initiative and was reduced for political reasons. Domestic SWHs used to be installed at a rate of 1000 m² (\$220,000) per year. No current statistics are available. Zimbabwe's political situation impacts the whole SADC.

SEE ALSO THE FOLLOWING ARTICLES

Biomass for Renewable Energy and Fuels • Computer Modeling of Renewable Power Systems • Development and Energy, Overview • Energy Ladder in Developing Nations • Potential for Energy Efficiency: Developing Nations • Renewable Energy and the City • Renewable Energy in Europe • Renewable Energy in the United States • Renewable Energy Policies and Barriers • Renewable Energy, Taxonomic Overview • Renewable Portfolio Standard

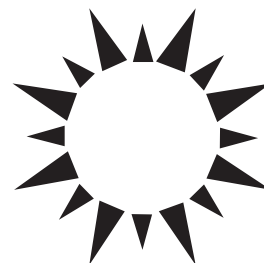
Further Reading

- Development Bank of Southern Africa/International Solar Energy Society (2000). "Renewable Energy Technologies in Southern Africa—A Guide for Investors." Midrand, Freiburg, Germany.
- Makenzie-Hoy, T. (2003). The DRC hydropower dream deconstructed. *Eng. News*, August 1–7, 56.
- Mohamed, S. (2003). Nepad and global governance. *Eng. News*, August 1–7, 54.
- Powering Nepad (2003). *Eng. News*. August. Available: <http://www.engineerignews.co.za/eng/sector/energy/?show=38166>.
- Schramm, G. (1998). Electrification programmes and the role of international development banks. In "ISES Utility Initiative for Africa" (D. Holm and W. Berger, Eds.), p. 107.
- Shell International (2002). "Shell Scenario." Shell International, London.
- Solarex Corporation. (1992). World design insolation [map]. Solarex Corp., Frederick, MD.
- Southern African Development Community (2002). SADC review. Available: <http://www.sadcreview.com>.
- Sustainable Energy Society of Southern Africa (2002). "Comments to the Draft White Paper on the Promotion of Renewable Energy and Clean Development." Sustainable Energy Society of Southern Africa, La Montagne, South Africa.

Renewable Energy in the United States

JOHN CARLIN

United States Department of Energy
Washington, D.C., United States



-
1. Introduction
 2. Overview of Renewable Energy
 3. Biomass Energy
 4. Hydroelectric Energy (Hydropower)
 5. Geothermal Energy
 6. Wind Energy
 7. Solar Energy

Glossary

biomass integrated gasification combined cycle A composite system that converts biomass feedstock into gas fuel for an electric generating unit that consists of one or more gas turbines with a portion of the required energy input provided by exhaust heat of the turbine to increase efficiency.

closed-loop biomass A process in which biomass feedstocks are grown specifically for the purpose of energy production. During the growing process, the feedstock absorbs CO₂, offsetting the forthcoming CO₂ emissions to be realized during energy use.

cofiring generation The use of a supplemental fuel in a boiler in addition to the primary fuel the boiler was originally designed to use.

criteria pollutant A pollutant determined to be hazardous to human health and regulated under the Environmental Protection Agency's (EPA) National Ambient Air Quality Standards. The 1970 amendments to the Clean Air Act require the EPA to describe the health and welfare impacts of a pollutant as the "criteria" for inclusion in the regulatory regime.

district heat Steam or hot water from an outside source used as an energy source in a building. The steam or hot water is produced in a central plant and piped into the building. The district heat may be purchased from a utility or provided by a physical plant in a separate building that is part of the same facility (e.g., a hospital complex or university).

greenhouse gases Those gases, such as water vapor, carbon dioxide, nitrous oxide, methane, hydrofluorocarbons, perfluorocarbons, and sulfur hexafluoride, that are transparent to solar (short-wave) radiation but opaque to long-wave radiation, thus preventing long-wave radiant energy from leaving the earth's atmosphere. The net effect is a trapping of absorbed radiation and a tendency to warm the planet's surface.

hot dry rock Heat energy residing in impermeable, crystalline rock. Hydraulic fracturing may be used to create permeability to enable circulation of water and removal of the heat.

landfill gas Natural gas (primarily methane gas) produced as a result of decaying waste at municipal landfills.

one sun Natural solar insulation falling on an object without concentration or diffusion of the solar rays.

renewable energy Resources that are naturally replenishing but flow limited. They are virtually inexhaustible in duration but limited in the amount of energy that is available per unit of time. Renewable energy resources include biomass, hydroelectric, geothermal, wind, and solar. Alternate transportation fuels include liquefied petroleum gas, ethanol, methanol, natural gas, electricity, and biodiesel.

run-of-river hydroelectric plant An electric generating plant using the power of moving water as it occurs and having little or no reservoir capacity for storage.

Interest in the use of renewable energy in the United States has increased during the past few decades beginning with the oil embargo of the early 1970s. Renewable energy is used in the residential, commercial, industrial, and transportation sectors. It is used to produce electricity at central grid-connected generators and at dispersed locations; cogeneration primarily in the industrial sector; heat in the industrial, residential, and commercial sectors; and ethanol for vehicle propulsion. During this time period, the U.S. energy policy has evolved around

two central themes: (i) energy/economic security and (ii) resource and environmental sustainability.

1. INTRODUCTION

Energy/economic security refers to the reliability of energy supplies and the stability of energy prices, which significantly influence individual and business expenditures and investment decisions. In the past, disruptive energy supplies and volatile energy prices have detrimentally affected economic growth and made decisions to invest in additional energy supplies and new energy technologies more difficult and risky. Renewable energy has the potential to mitigate these effects. Its prices are not volatile, and once the capital is invested, the fuel prices are stable and in most cases negligible. Any reduction in the demand for conventional fuels reduces the price of that conventional fuel for all units consumed. Even if a new technology does not actually penetrate the market, its availability, on the shelf, limits price levels for conventional fuels. Renewable energy also enhances energy diversification, reduces capital outflows, and utilizes domestic labor and resources.

Renewable energy also plays an important role in reducing emissions from fossil fuels when used in their place. These emissions include greenhouse gases and criteria pollutants such as sulfur oxide (SO_x) and nitrous oxide (NO_x). Renewable fuels, with the exception of biomass, have little or no emissions. The use of closed-loop biomass reduces carbon emissions, a very important component of greenhouse gases. Cofiring of biomass and coal to produce electricity may also reduce the emission of SO_x and NO_x .

The use of renewable energy also has some disadvantages, principally related to cost. Renewable technologies are usually more expensive than conventional technologies and frequently require larger up-front capital investments. Also, there are land-use and other environmental issues related to their use.

The following are the most important renewable fuels in use today:

Hydroelectric power (hydropower): The use of flowing water to produce electrical energy.

Biomass/biofuels: Organic matter that includes forest thinnings, forest and mill residues, agricultural by-products and crops, derived oils and wastes, wood and wood wastes, animal wastes, livestock operation residues, aquatic plants, fast-growing trees and plants, and municipal and industrial

wastes. Biofuels refers to liquid fuels derived from biomass feedstocks.

Geothermal energy: Hot water or steam extracted from geothermal reservoirs in the earth's crust that can be used for geothermal heat pumps, space and district heating, water heating, or electricity generation.

Wind energy: Kinetic energy present in wind motion that can be converted to mechanical energy for driving pumps, mills, and electric power generators.

Solar energy: The radiant energy of the sun, which can be converted into other forms of energy, such as heat or electricity.

This article presents basic information on renewable energy including a cursory discussion of its principal elements. The focus is on providing data for 2001 while explaining the role of renewable technologies in the U.S. energy marketplace.

2. OVERVIEW OF RENEWABLE ENERGY

The use of renewable energy in the United States peaked during the mid-1990s at approximately 7.5 quadrillion Btus (quads), increasing from slightly more than 4 quads in 1970 (Fig. 1). Almost half of this increase occurred in the electric power sector and another one-fourth occurred in the industrial sector. The remaining growth occurred in the residential, commercial, and transportation sectors and net imports of renewable energy. Hydropower, biomass, and geothermal energy account for most of the growth.

Since the mid-1990s, however, renewable energy consumption has declined approximately 25% due to the reduced water availability for hydropower. In 2001, renewable energy consumption amounted to 5.668 quadrillion Btus (Fig. 2), which represents 6% of total U.S. energy consumption. Even with the decline in the availability of hydropower, biomass and hydropower still accounted for 92% of total renewable energy consumption.

In 2001, more than 60% of total renewable energy consumed (3.5 quads) was used to produce electricity (Table I). In addition, the industrial sector used 1.4 quads for nonelectric purposes (process heat and steam). The renewable energy use in the industrial sector was primarily a wood or wood waste feedstock. The transportation sector consumed only 0.1 quads of renewable energy

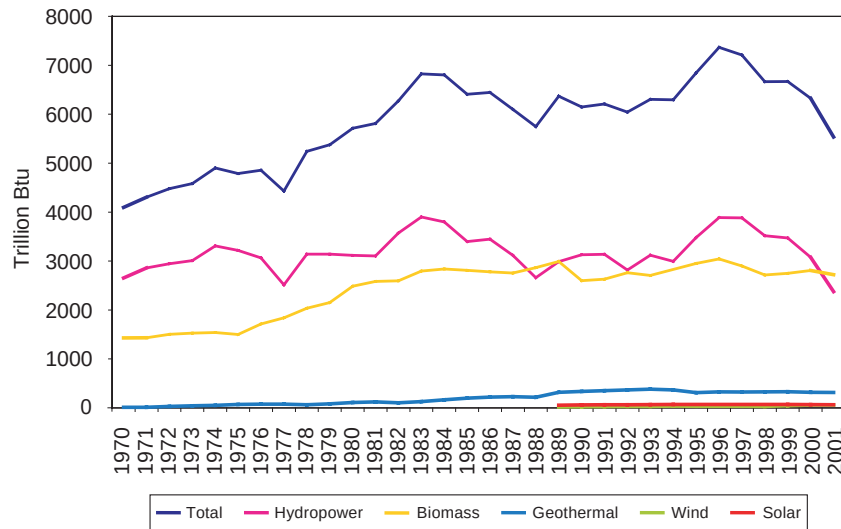


FIGURE 1 Growth in renewable energy consumption by source. From Energy Information Administration (2002). Annual energy review, 2001. Energy Information Administration, Washington, DC.

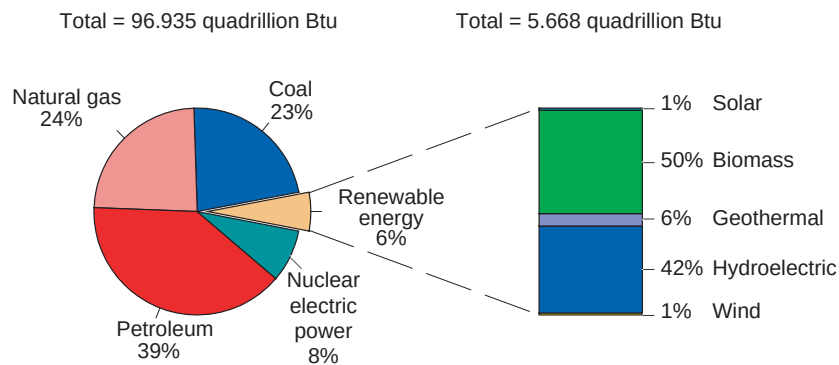


FIGURE 2 The role of renewable energy consumption in the nation's energy supply, 2001. From Energy Information Administration (2002, Fig. H1).

(ethanol). Although the use of ethanol in gasoline has doubled during the past decade, it represents only approximately 0.5% of total transportation fuel consumption.

The increased use of renewable technologies, other than hydropower, during the latter quarter of the 20th century is largely the result of government incentives and regulations (Table II). The federal government provided financial incentives and funds for research and development (R&D) during this period. By 1982, most renewable energy projects were eligible for a 10% investment tax credit, a 15% business renewable energy investment tax credit, a 40% residential tax credit, and a 5-year accelerated depreciation schedule. However, the most significant impetus to the evolution of the renewable industry

was the passage of the Public Utility Regulatory Policies Act of 1978 [Public Law (P.L.) 96-917].

3. BIOMASS ENERGY

The Energy Information Administration's (EIA) current data accounts divide biomass/biofuels into four categories: wood and wood waste, municipal solid waste, other biomass waste, and corn ethanol.

Wood and wood waste come primarily from U.S. forestlands or timberlands. Other sources include private land clearing and silviculture, and urban tree and landscape residues. Wood is almost entirely derived from U.S. timberlands, which contain roundwood, timber stock consisting of 92% sound trees

TABLE I

U.S. Renewable Energy Balances, 2001^a

Sector and source	Energy consumption (quadrillion Btus)			Electricity net generation (thousand kWh)	Electricity net summer capacity (MW)
	Electric	Nonelectric	Total		
Total	3.541	2.127	5.668	312,767,216	96,741
Biomass	0.808	2.046	2.854	59,640,051	10,120
Geothermal	0.292	0.022	0.315	13,901,229	2793
Conventional hydroelectric	2.376	—	2.376	232,949,965	79,379
Solar	0.005	0.059	0.064	494,158	387
Wind	0.059	—	0.059	5,781,813	4062
Residential		0.475	0.475		
Biomass		0.407	0.407		
Geothermal		0.009	0.009		
Solar		0.059	0.059		
Commercial	0.027	0.072	0.098	2,063,254	433
Biomass	0.026	0.063	0.089	1,963,505	399
Geothermal		0.008	0.008		
Conventional hydroelectric	0.001		0.001	99,749	35
Industrial	0.402	1.414	1.816	32,361,740	5667
Biomass	0.365	1.409	1.774	28,739,925	4588
Geothermal		0.005	0.005		—
Conventional hydroelectric	0.037		0.037	3,621,815	1079
Transportation					
Biomass (alcohol fuels)		0.133	0.133		
Electric power	2.953	0.034	2.987	262,853,221	90,640
Biomass	0.417	0.034	0.451	28,936,621	5134
Geothermal	0.290		0.290	13,812,908	2793
Conventional hydroelectric	2.181		2.181	213,827,721	78,265
Solar	0.005		0.005	494,158	387
Wind	0.059		0.059	5,781,813	4062
Net renewable energy imports	0.159		0.159	15,489,001	
Geothermal	0.002		0.002	88,321	
Conventional hydroelectric	0.241		0.241	23,610,560	
Conventional hydroelectric (exports)	0.084		0.084	8,209,880	

^a Source. Energy Information Administration. (2002).

(called growing stock) and 8% rotten, cull, or “salvable” dead trees.

Pure wood for fuel (known as fuelwood) actually accounts for a very small percentage of biomass energy consumption. The majority is wood waste. Waste wood includes principally manufacturing and wood processing wastes from wood mills. Primary

wood mills include sawmills, veneer mills, and pulp mills. Secondary mills include manufacturers of dimension lumber, trusses and building components, flooring, windows and doors, cabinets, pallets, poles and fencing, barrels, boats, highway transport trailers, manufactured homes, and musical instruments. Mill waste residues are used primarily as fuel,

TABLE II

Key Milestones for Renewable Fuels

Year	Milestone	Description
1973	OPEC oil embargo	OPEC disrupted oil supplies and raised crude oil prices by 70%.
1974	Oil embargo ends	Project independence was Initiated to review strategic energy options.
1978	Public Utility Regulatory Policies Act of 1978 (PURPA)	PURPA mandated the purchase of electricity at the utilities "avoided cost" from qualifying facilities (QFs) meeting certain technical standards regarding energy source and efficiency. QFs were also exempted from both state and federal regulation under the Federal Power Act and the Public Utility Holding Company Act.
1978	Energy tax credit	A 15% energy tax credit was added to an existing 10% investment tax credit, providing incentive for capital investment in photovoltaic applications and solar thermal generation facilities for independent power producers.
1978	Energy tax act	The Energy Tax Act of 1978 (Public Law 95-618) gave a ¢ per gallon exemption from federal excise taxes to motor fuels blended with ethanol (minimum 10% ethanol) and granted a 10% energy investment tax credit for biomass-ethanol conversion equipment (in addition to the 10% investment tax credit available).
1980	Crude Oil Windfall Profits Act	The act (Public Law 96-223) extended the per gallon federal excise tax exemption for ethanol blends to December 31, 1992, and extended the energy investment tax credit to December 31, 1985. An income tax credit was also provided to alcohol fuel blenders—40¢ per gallon for 190-proof alcohol and 30¢ per gallon for 150–190 proof. The total federal tax credit for a turbine was increased to 25%. The act also raised the residential tax credit to 40% of the first \$10,000 investment in photovoltaic applications.
1980	Energy Security Act	The act (Public Law 96-294) offered insured loans for small ethanol producers (less than 1 million gallons per year), loan guarantees that covered up to 90% of construction costs on ethanol plants, price guarantees for biomass energy projects, and purchase agreements for biomass energy used by federal agencies. It also established the Department of Energy (DEO) Office of Alcohol Fuels and authorized \$600 million for both the U.S. Department of Agriculture (USDA) and the DOE for biomass research. Subsequent rescissions altered this allocation to \$20 million for the USDA and \$800 million for the DOE to use for alcohol fuel loans. The Consolidated Farm and Rural Development Act of 1980, which rescinded the \$505 million allocated to the USDA, appropriated \$250 million for alcohol loan guarantees that were used to support 12 firms.
1981	California State energy tax credit	The state of California enacted a 25% tax credit for the capital costs of renewable energy systems.
1983	California Standard Offer Contracts	California's Standard Offer Contract system for QFs provided renewable electric energy systems a relatively firm and stable market for their output.
1984	Tax Reform Act	The act (Public Law 97-424) raised the gasohol exemption from 5 to 6¢ per gallon, with the overall tax unchanged at 9¢ per gallon of retail fuel. The blender's income tax credit was increased to 60¢ per gallon for the 190-proof alcohol and 45¢ for the 150–190 proof.
1990	Omnibus Budget Reconciliation Act	The act (Public Law 101-508) decreased the gasohol tax exemption from 6 to 5.4¢ per gallon. Tax credits for neat

continues

Table II continued

Year	Milestone	Description
1990	Clean Air Act Amendments	ethanol sales remained unchanged at 6¢ per gallon. The expiration date was extended to 2002. The amendments (Public Law 101-549) mandated the winter use of oxygenated fuels in 39 major carbon monoxide nonattainment areas (areas where Environmental Protection Agency emissions standards for carbon dioxide had not been met) and required year-round use of oxygenates in 9 severe ozone nonattainment areas in 1995. The 1990 amendments also establish the Acid Rain Program, which is designed to reduce the adverse effects of acid deposition. This improvement will be achieved primarily through reductions of SO ₂ and NO _x emissions by electricity producers.
1992	Energy Policy Act	The act (EPACT) (Public Law 102-486) established a permanent 10% tax credit for investments in solar and geothermal equipment. Established a 10-year, 1.5¢ per kilowatt-hour (kWh) production tax credit for privately owned as well as investor-owned wind projects and biomass plants using dedicated crops (closed loop) brought on line between 1994 and 1993, respectively, and June 30, 1999. The act instituted the Renewable Energy Production Incentive, which provides a 1.5¢/kWh incentive, subject to annual congressional appropriations (section 1212), for generation from biomass (except municipal solid waste), geothermal (except dry steam), wind, and solar from tax exempt publicly owned utilities and rural cooperatives. Indefinitely extended the 10% business energy tax credit for solar and geothermal projects. It provides a tax credit (variable by gross vehicle weight) for dedicated alcohol-fueled vehicles, a limited tax credit for alcohol dual-fueled vehicles, and, a tax deduction for alcohol fuel dispensing equipment.
1995	Federal Energy Regulatory Commission (FERC) prohibition on QF contracts above avoided cost	In a ruling against the California Public Utility Commission, FERC refused to allow a bidding procedure that would have the effect of allowing rates above avoided cost from renewable QFs.
Mid-1990s	ISO4 contract Rollover in California at lower rates	Ten-year QF contracts written during the mid-1980s at rates of 6¢/kWh and higher began rolling over at mid-1990s avoided costs of approximately 3¢/kWh. This "11th-year cliff" creates financial hardship for most QFs on ISO4 contracts.

with the next most important application being use as pulp and fiber for making paper products. The paper production process, in turn, yields a by-product known as black liquor, which is the single largest source of biomass energy. Pulp and paper mills use large quantities of wood bark, edgings, and residues from their own log-stripping operations and also buy them from facilities such as sawmills.

Waste wood also includes construction and demolition debris, wood from pallets and containers, and wood pellets. Waste wood is also salvaged for fuel by private owners from woodlots, farm fence rows, cropland clearing, orchards, and operations of private urban silviculture. Urban tree and landscape residues consist of tree limbs, tops, brush, leaves, stumps, and grass clippings. They are generated by

commercial tree care firms, municipal tree trimming operations, electric utility power line maintenance departments, orchards, and landscapers.

Municipal solid waste (MSW) includes durable and nondurable goods; containers and packaging; food wastes; yard wastes; and miscellaneous waste from residential, commercial, institutional, and some industrial sources. Examples of MSW include yard trimmings (lawn clippings and branches), newspapers and magazines, office paper, boxes and packaging, glass, metal, plastics, rubber, clothing, and food scraps and cafeteria wastes. MSW also includes such large and durable items as furniture, wood pallets, and appliances. MSW management begins with methods of reducing the quantities of trash generated in the first place (source

reduction). However, once the trash has been produced, MSW management involves four components: composting, recycling, waste-to-energy via combustion, and landfilling.

The last two categories of biomass/biofuels are relatively small. Other biomass waste is a catch-all category that includes agricultural by-products and crops; sludge waste; tires; and other biomass solids, liquids, and gases. Biofuels consists of liquid biomass (primarily ethanol from corn) used in the transportation sector.

3.1 Current Data

In 2001, biomass/biofuels contributed more than 2.8 quads to U.S. energy accounts (Table III). More than 2 quads of this energy was used to produce steam or process heat for the industrial sector. The remaining 0.8 quads was used to produce approximately 60 billion kilowatt-hours (kWh) of electricity generation.

More than half of the electricity generation came from the electric power sector, with most of the remainder resulting from cogeneration in the industrial sector. There was a small amount of cofiring with coal in the electric power sector.

The largest fuel category was wood/wood waste, comprising more than 75% of the total account balance. Nine products represent the wood/wood waste category: black liquor, red liquor, sludge wood, spent sulfite liquor, peat, paper pellets, railroad ties, utility poles, and wood/wood waste. The second largest contributor was the use of MSW at waste-to-energy combusting facilities and landfills. Other biomass, which consists of agricultural by-products/crops, straw, sludge waste, tires, medical wastes, solid by-products, fish oil, liquid acetonitrile waste, tall oil, waste alcohol, digester gas, and methane, amounted to 117 trillion Btus. The transportation sector, consisting of ethanol from corn, contributed only 133 trillion Btus to the data accounts.

TABLE III

U.S. Biomass/Biofuels Energy Balances, 2001^a

Sector and source	Energy consumption (quadrillion Btus)			Electricity net generation (thousand kWh)	Electricity net summer capacity (MW)
	Electric	Nonelectric	Total		
Biomass	0.808	2.046	2.854	59,640,051	10,120
Alcohol fuels	—	0.133	0.133		
Wood/wood waste	0.483	1.687	2.170	36,871,734	6230
MSW/landfill gas	0.292	0.142	0.434	20,018,830	3387
Other biomass	0.033	0.084	0.117	2,749,487	504
Residential					
Biomass		0.407	0.407		
Wood		0.407	0.407		
Commercial					
Biomass	0.026	0.063	0.089	1,963,505	399
Wood/wood waste	0.000	0.043	0.043	19,523	7
MSW/landfill gas	0.021	0.019	0.040	1,539,085	299
Other biomass	0.005	0.001	0.006	404,897	96
Industrial					
Biomass	0.365	1.409	1.774	28,739,925	4588
Wood/wood waste	0.354	1.227	1.580	27,735,132	4388
MSW/landfill gas	0.001	0.103	0.104	61,286	21
Other biomass	0.011	0.079	0.090	943,507	179
Transportation					
Biomass (alcohol fuels)		0.133	0.133		
Biomass	0.417	0.034	0.451	28,936,621	5134
Wood/wood waste	0.130	0.011	0.140	9,117,079	1835
MSW/landfill gas	0.271	0.020	0.290	18,418,459	3070
Other biomass	0.017	0.004	0.021	1,401,083	229

^aSource. Energy Information Administration (2002).

3.2 Technology and Applications

Biomass combustion in the electric power, industrial, and residential sectors is the principal energy use of biomass today. To a much lesser extent, biomass can be transformed into a gas such as landfill gas, and biofuels can be created for use in the transportation sector. Biomass-fired combustion systems are commercially available to produce steam, electricity, and heat in various applications in the industrial, electric power, and residential sectors. Direct combustion systems in the industrial and electric power sectors are usually classified according to the way the biomass is burned: pile, grate, suspension, or fluidized bed.

A Dutch oven is an example of a pile-burning system. Two chambers are used for combustion: The first chamber dries and partially burns the fuel, and the second chamber completes the burning process. Some versions of the Dutch oven design include grates that facilitate drying and transport of the fuel and cofiring capability. The spreader-stoker is an example of a grate-burning system. After the fuel is spread on a grate, it is blown into a combustion chamber and burnt in suspension. Any unburnt fuel falls onto a grate, where combustion is completed. In suspension burners the fuel is burnt in an air stream. Examples of suspension burners are cyclonic burners and solid fuel injection burners. A heated bed of sand kept in constant motion is the surface for fuel combustion in a fluidized-bed combustion system.

An advanced biomass combustion technology is the biomass integrated gasification combined-cycle generating plant. It is assumed that hot gas filtration will be used for gas cleanup purposes in this technology. The advantages of the newer hot gas cleanup technology are that it generates less wastewater and it allows the process to be operated at higher efficiencies than those of the cold gas cleanup alternative. The disadvantages are limited operational experience, higher costs, and greater complexity.

MSW is combusted into energy throughout the country, producing steam, electricity, or both. Two basic approaches are used: mass burning and refuse-derived fuel (RDF). The two most common types of mass-burn facilities are waterwall incinerators and the smaller modular combustion units. At mass-burn facilities, residential and commercial waste is deposited into a pit and moved to the furnaces with overhead cranes with little sizing or separation. The wastes are usually burned on a grate and the heat is used to create steam, which can be sold directly to industrial or commercial customers or used to power

a turbine to generate electricity. Waterwall furnace designs burn waste, heating the water that circulates through the walls of the combustion chamber in steel tubes, resulting in the production of steam. Modular mass-burn facilities are small-scale combustion units that can be fabricated at the factory and shipped fully assembled or shipped in modules and assembled at the construction site.

The first stage of a RDF system is the preparation of the fuel. Raw MSW is processed to form a more homogeneous product by removing noncombustible materials, such as metal, glass, grit, and sand. Sometimes, the remaining RDF is pelletized. It may be sold to outside parties or burned on-site in a furnace. In some cases, RDF is cofired with coal.

The residential/commercial sector uses five types of wood-burning equipment: free-standing stoves, fireplaces, fireplace inserts, central heating equipment for houses, and large boilers for apartment buildings and commercial buildings. Each type is produced in different sizes, has different performance parameters, and is made from a variety of construction materials (steel, iron, and alloys). Fireplaces are the most common type but are rarely used for heating an entire house because of their low efficiency and rapid consumption of wood. Many fireplaces result in a net heat loss because they require a large amount of air from the house that is replaced with cold outside air. Improvements to fireplaces can be made by adding enclosures, heat blowers, an outside air source, or fireplace inserts. Free-standing stoves are the main type of wood-burning equipment used for home heating. The two basic woodstove designs are the radiating and the circulating stoves.

Biomass (i.e., MSW in landfills) is also a source of landfill gas (LFG) that can be recovered and the energy can be used for direct heating, electricity generation, chemical feedstock, purification to pipeline-quality gas, and heat recovery. Each of these methods has a variety of LFG applications:

1. Direct heating applications
 - Industrial boilers
 - Space heating and cooling
 - Industrial heating/cofiring
2. Electricity generation applications
 - Processing and use in reciprocating internal combustion engines
 - Processing and use in gas and steam turbines
 - Processing and use of fuel cells

3. Feedstock in chemical manufacturing processes
 - Conversion to methanol (and optional subsequent industrial or vehicular fuel use)
 - Conversion to diesel fuel (and subsequent use as vehicular fuel)
4. Purification to pipeline-quality gas
 - Utilization as vehicular fuel
 - Incorporation into local natural gas network
5. Heat recovery from landfill flares

The primary approach for converting biomass to alcohol fuels is the fermentation of corn into ethanol. The largest potential source of ethanol in the United States is the biochemical conversion of cellulosic biomass into ethanol.

3.3 Niche Markets and Outlook

The conversion of biomass into energy is in large part a waste disposal option. The waste is already available, close by, and must be disposed of. In these cases, the amount of waste available is a constraint to growth in the biomass energy industry. In addition, there are several existing and potential niche markets for biomass energy as a means of alleviating environmental and security concerns associated with the use of fossil energy.

There are approximately 30 facilities that cofire biomass and coal in existing coal-fired boilers. Coal-fired boilers can combust a premixed combination of biomass and coal fed through an existing coal-feed system or retrofitted with a separate feed system for biomass. Generally, the proportion of biomass consumed ranges from 1 to 10%. The substitution of biomass for coal at electric power plants reduces emissions of SO_2 and NO_x .

A second niche market that is growing very rapidly is the conversion of LFG into energy. Several factors have contributed to this growth: U.S. Supreme Court rulings that eliminated flow control contracts that previously directed the waste to in-state waste-to-energy combustion facilities, federal tax credits, and environmental benefits. The environmental advantages of energy recovery include decreased emissions of methane, nonmethane organic compounds, and toxics (e.g., benzene, carbon tetrachloride, and chloroform). Although carbon dioxide (CO_2) emissions increase with the energy recovery option, the net atmospheric balance is a positive one because CO_2 emissions are significantly less radiative

(i.e., the alleged greenhouse effect is less) than methane emissions.

Another possible niche market application for biomass is in the transportation sector. As mentioned previously, ethanol is the only renewable fuel used in substantial quantities in the transportation sector. Ethanol is made from corn. During the past two decades, it has received an excise tax exemption, generally 5 or 6¢ per gallon on a 10% ethanol/90% gasoline blend. Even with a subsidy, methyl tertiary-butyl ether was viewed until recently as the more economic and practical choice as an oxygenate additive to gasoline or as a component of RFG.

Ethanol can also be the primary component in a transportation fuel, such as E85 (85% ethanol/15% gasoline). Despite the fact that there are approximately 3 million vehicles on the road that can use any combination of E85 and regular gasoline, only 0.7% of all ethanol is E85. This reflects the difficulty in developing an E85 refueling infrastructure, and the fact that E85 on a heating value basis has generally been at least as costly as gasoline. Should a less costly source/method of ethanol become available (e.g., lignocellulosic biomass conversion and biomass gasification), fleets capable of centralized refueling would provide a major entry point for E85 into the transportation sector.

For biomass to provide large amounts of energy, dedicated crops must be developed and used for electricity generation and as an alternative to gasoline in the transportation sector. To the extent that biomass is closed-loop, its use as a substitute for fossil fuel results in a reduction in CO_2 emissions, the most significant and most difficult to harness of the greenhouse gases. The use of biomass as a substitute for gasoline could potentially mitigate energy/economic concerns associated with the import of oil from the Organization of Petroleum Exporting Countries.

4. HYDROELECTRIC ENERGY (HYDROPOWER)

Water resources that are used for power production are found in 48 of the 50 states. Generating electricity using water has several advantages. The major advantage is that water is a source of relatively inexpensive power. In addition, because there is no fuel combustion, there is little air pollution in comparison with fossil fuel plants and limited thermal pollution compared with nuclear plants. Like other energy sources, the use of water for

generation has limitations, including environmental impacts caused by damming rivers and streams, which affects the habitats of the local plant, fish, and animal life.

4.1 Current Data

As of 2001, conventional hydroelectric net summer capacity totaled 79,379 MW, which generated 232,949,965 thousand net kWh of electricity. Of that total, 92% was generated by electric utilities and independent power producers in the electric power sector, followed by generation in the industrial and commercial sectors, respectively.

4.2 Technology and Applications

There are two types of conventional hydroelectric plants: storage and run-of-river. Storage plants consist of a dam, which impedes water and regulates its flow. These plants store up to a season's supply of water. Run-of-river plants have very small impoundment, storing a day's supply of water. Source water may be from free-flowing rivers, streams, or canals. In both types of plants, the water is guided to a turbine by a penstock (tube), which in turn drives a generator to produce electricity. The amount of generation is dependent on precipitation. This is especially true for run-of-river plants, which have very small impoundment. Hydroelectric facilities can be used when demand is high because hydro units can respond to rapid load changes, which is not possible with steam-powered thermal facilities.

Challenges faced by the industry include mandatory relicensing for investor-owned facilities, improving operating efficiency, and mitigating adverse environmental effects. The U.S. Department of Energy, through its Advanced Hydropower Turbine System program, is working to improve operating efficiency and will assist in addressing some of the environmental issues. The program is expected to be completed by 2010.

4.3 Niche Markets and Outlook

The relicensing process is placing new conditions on hydroelectric operations and, therefore, productivity. Generally, relicensed plants (i.e., the current stock) will have slightly decreased generation. Although most good sites have already been developed, it was estimated in 1998 that there were 5677 sites with a total undeveloped capacity of approximately 30,000 MW available for power production, after

accounting for environmental, legal, and institutional constraints. This total includes approximately 8500 MW where no facility exists, approximately 17,000 MW where a dam exists but no hydroelectricity is generated, and approximately 4300 MW of incremental hydropower potential at existing sites from capacity additions or improved efficiencies. Finally, there is some, as yet not quantified, amount of capacity that could be added by refurbishing existing plants.

5. GEOTHERMAL ENERGY

Geothermal energy is heat in the earth's crust brought to the surface from underground reservoirs of steam or hot water. As used at electric generating facilities, hot water or steam (hydrothermal resources) extracted from the earth's crust supplies steam to turbines that drive generators to produce electricity at electric power plants. Eventually, it is hoped that continued development of extraction methods will allow geothermal power production from so-called "hot dry rock" formations, which would open significantly more resources than currently utilized. Most of the known high-temperature hydrothermal resource areas in the United States are in the western part of the country. Electricity from geothermal resources is currently produced in California, Nevada, Utah, Hawaii, and New Mexico.

5.1 Current Data

In 2001, there were 43 operational geothermal power plants in the United States. Those with a capacity of 1 MW or more totaled 2793 MW, which generated 13,813 thousand MWh of electricity, the third most of all renewables after hydroelectricity and biomass. Direct use applications included space and district heating; industrial, such as food dehydration (the drying of vegetable and food products); greenhouse; aquaculture; spas and pools; and snow melting. These contributed 9 trillion Btus to U.S. energy consumption in 2001, whereas geothermal heat pumps contributed 13.5 trillion Btus. The energy consumption of each roughly doubled since 1990. Shipments of geothermal heat pumps by manufacturers have averaged 36,869 per year during the period 1996–2000.

With regard to environmental considerations, geothermal energy does not require fuel burning to produce heat or electricity. Although geothermal steam and flash plants emit mostly water vapor, some

harmful compounds, such as sulfur oxides, nitrogen oxides, hydrogen sulfide, and particulates, are also released. These plants use available emission control systems to meet the most stringent clean air standards. Binary cycle power plants run on a closed-loop system, so no gases are emitted. Whereas binary cycle plants are typically air cooled, the other plant types require a source of cooling water.

Other land uses can coexist with geothermal plants with little interference or fear of accidents. At some sites, however, there are important cultural objections to geothermal development from native populations.

Finally, a typical geothermal plant requires several wells. Although drilling these wells has an impact on the land, using advanced directional or slant drilling minimizes the impact. Several wells can be drilled from one pad, so less land is needed for access roads and fluid piping.

5.2 Technology and Applications

5.2.1 Power Plants

Geothermal power plants are reliable and productive, with a secure fuel supply. They provide baseload power and operate at capacity factors of 85–90% or more. Plants can be built in a modular fashion. This modularity provides flexibility; additional units can be added as needed.

There are three types of geothermal electricity-generating technologies in operation: dry steam, flash steam, and binary cycle. Dry steam plants use steam directly to turn a turbine, which drives a generator that produces electricity. In the United States, these plants are found exclusively at geysers, a high-quality steam-dominated resource.

Hydrothermal fluids above approximately 338°F (170°C) can be used in flash steam plants to make electricity. For these power plants, hot geothermal fluids come out of the ground under pressure and above boiling, and are sprayed into a tank held at a much lower pressure than the fluid, causing some of the fluid to rapidly vaporize or “flash” to steam. The steam then drives a turbine and generator. A second-stage, even lower pressure phase is employed in dual-flash plants to extract additional energy from fluids unflashed in the first phase.

Most geothermal areas contain moderate-temperature water (<338°F). Energy is extracted from these fluids in binary cycle power plants. Hot geothermal fluid and a secondary (hence, binary) fluid with a much lower boiling point than that of water pass through a heat exchanger. Heat from the

geothermal fluid causes the secondary fluid to vaporize and drive the turbines.

5.2.2 Direct Use Applications

Direct use systems use lower temperature water [68–302°F (20–150°C)] to provide direct heat for residential, industrial, and commercial uses. These uses of heat from the earth eliminate the need to burn fuel, preventing thousands of tons of greenhouse gases from being emitted.

Direct use systems are typically composed of three components: a production facility, usually a well, to bring the hot water to the surface; a mechanical system—piping, heat exchanger, and controls—to deliver the heat to the space or process; and a disposal system— injection well or storage pond—to receive the cooled geothermal fluid.

The geothermal heat pump is an electric heat pump that uses the earth as a heat source for heating or as a heat sink for cooling. A water and antifreeze mixture circulates through pipes buried in the ground or in subsurface waters (a vertical or horizontal closed loop) and transfers thermal energy to a heat exchanger in the heat pump. The heat exchanger works through a water-to-refrigerant loop. Another variation is the open-loop system that uses groundwater pumped to the heat exchanger in the heat pump. Geothermal heat pumps offer a distinct advantage over the use of air as a source or sink because the ground is at a more favorable temperature. Compared to air, the ground is warmer in winter and cooler in summer. Therefore, geothermal heat pumps demonstrate better performance over air-source heat pumps.

5.3 Niche Markets and Outlook

Challenges faced by the geothermal power industry include the significant risk and uncertainty in locating and tapping commercially useful resources, even in land areas known to have geothermal resources; the high cost of drilling wells; the lack of advanced technology to find and characterize geothermal resources without expensive drilling; and the need to develop technology to allow commercial production from beyond the limited natural hydrothermal resource base.

Current geothermal energy research is addressing these challenges by focusing on resource exploration, drilling technologies, reservoir engineering, and power plant operating efficiencies. The objective of this research is to find and utilize available resources and lower the cost of producing electric power.

6. WIND ENERGY

Winds are created by uneven heating of the atmosphere by the sun, irregularities in the earth's surface, and the rotation of the earth. As a result, winds are strongly influenced and modified by local terrain, bodies of water, weather patterns, vegetative cover, and other such factors. Wind turbines work because they slow down the speed of the wind. The wind flows over the airfoil-shaped blades creating lift, similar to the effect on airplane wings, causing them to turn. The blades are connected to a driveshaft that turns an electric generator to produce electricity.

Although wind energy currently provides less than 2% of U.S. renewable electricity generation, the United States is rich in wind resources. However, the small fraction of this resource that can be economically exploited and the quantity of electricity that can be accommodated on the transmission grid remain highly uncertain in the evaluation of the total resource potential. Some states with good wind resources have difficult terrain, large areas excluded from development for environmental or other land use reasons, or, as in the case of North and South Dakota, are far from demand centers.

Wind power development in the United States began in the early 1980s and was almost exclusively in California due to the favorable regulatory environment and incentives for renewable energy development. It was not until 1999 that significant wind capacity appeared in other states, such as Iowa, Minnesota, and Texas. Development in these states was due in part to state mandates and renewable portfolio standards. In 2001, 20 states reported wind capacity, for a U.S. total of more than 4000 MW. The current expansion is possible because the leveled cost of wind power is becoming competitive with that of power from fossil fuels particularly when the federal production tax credit is factored in, coupled with a wide range of state incentives that benefit wind development. Excluding California, whose wind plants were developed mainly in the 1980s and 1990s, approximately 60% of U.S. wind capacity added during the 3-year period 1999–2001 was in states that currently have renewable portfolio standards.

6.1 Technology and Applications

The most common turbine used in the United States is the three-bladed horizontal axis upwind turbine, but two-bladed horizontal axis downwind turbines and vertical axis turbines also exist. Lattice- or tubular-

type towers support the rotor blades and nacelle (an enclosure that houses key components of the turbine, including the gearbox and generator). In 2000, turbines with 600- to 900-kW capacity were popular at new wind farms. However, larger wind farms have been developed since using models with 1300- to 1800-kW capacity. The larger turbines can usually take advantage of some economies of scale and faster wind speeds at the higher heights of their towers. Rotors for the 1500-kW turbines are typically 70–80 m in diameter and are mounted on towers 60–80 + m tall. In the future, multi-megawatt turbines may be possible. For example, GE Wind Energy started testing a prototype 3600-kW turbine in 2002. This model has been selected for the proposed Cape Cod, Massachusetts, project, which may be one of the first offshore projects in the United States.

A handful of manufacturers supply most of the U.S. market for utility-scale turbines. Of these, only one, GE Wind, is domestic; the remainder are European owned. GE Wind came into being in mid-2002 after the acquisition of some of Enron Wind's assets. It is a vertically integrated company providing a wide array of services from planning to operations and maintenance. It has turbine-manufacturing facilities in Tehachapi, California, as well as in Spain and Germany.

Vestas is a leading manufacturer of wind turbines worldwide and supplies U.S. markets with nacelles and blades manufactured in its home country, Denmark. Towers for U.S. projects are procured from suppliers in the United States. There are plans for a future Vestas manufacturing facility in Portland, Oregon. NEG Micon is another Danish turbine manufacturer supplying the U.S. market. It had a manufacturing facility in Champaign, Illinois, but this is currently operating only as a service and repair facility. Its nacelles come from Denmark, whereas blades and towers come from U.S. suppliers. Bonus Energy also supplies U.S. markets with turbines, nacelles, and blades manufactured in Denmark; some towers used in Bonus projects are manufactured in the United States. Nordex is a German manufacturer of wind turbines with interests in the United States. Other domestic companies supply small wind turbines, such as Atlantic Orient Corporation (in the 50-kW range) and Bergey Wind (<10 kW).

6.2 Current Data

By the end of 2001, there were utility-scale wind farms in 20 states. Six states accounted for more than 88% of U.S. capacity (Table IV). In 2001, these wind farms

TABLE IV*U.S. Wind Nameplate Capacity (MW), 2001^a*

Total	4079
California	1608
Texas	1060
Minnesota	300
Iowa	277
Oregon	173
Washington	166

^aSource. U.S. Department of Energy, Energy Information Administration, unpublished data.

produced approximately 6 billion kWh of electricity or approximately 2% of total renewable generation. Wind farms were operating in 10 more states by the end of 2002, with more planned for 2003.

6.3 Niche Markets and Outlook

The nature of wind as an energy source dictates that wind farms have to be located where the wind is occurring. This may be distant from load centers. Furthermore, the location for optimal gathering of wind may be in a terrain in which it is very difficult to erect a turbine tower. This has led to the development of semi-self-erecting towers. Also, bird kills is an issue, especially when towers were constructed with trestles that inadvertently served as nesting places. Newer, sleek, tubular towers do not provide such spaces.

In some areas of the country, the transmission grid is seriously congested and it is difficult and costly to construct new lines. Wind is intermittent and commonly operates with capacity factors under 35%, so backup power must be available when the wind is not blowing. Sometimes, this is accomplished relatively easily by coupling wind power with hydropower, as it is frequently done in California. However, sometimes the situation is more complicated. Also, most grid tariffs and operating procedures were not designed to consider intermittent generation and may not adequately accommodate its unique characteristics. Although new procedures for rate determination are the subject of debate (e.g., the Federal Energy Regulatory Commission's Standard Market Design), the ability to develop technology that could more reliably forecast future wind conditions is an issue. Also, economic large-scale storage is not commercially available.

Federal support for wind power has had gaps. Twice in recent years, the Federal Production Tax Credit (PTC) has expired. On both occasions, there was a period of uncertainty, and project plans were halted or de-

layed for a number of months until the credit was renewed. The PTC will expire again at the end of 2003.

The EIA's forecast to 2025 projects that wind power will increase in the United States. However, the rate of growth slows in the latter years of the forecast as current incentives and mandates expire and wind economics remain somewhat marginal. The EIA projects that total U.S. wind capacity will expand to 12,000 MW by 2025.

7. SOLAR ENERGY

Using the sun's energy to produce heat or electricity is not a new idea. Technologies that capture the radiation of the sun and use it to produce energy have been available in one form or another for centuries. They continue to be defined and developed today. As of 2003, solar thermal devices were used in a variety of ways, from heating swimming pools to creating steam for electricity generation, whereas photovoltaic (PV) devices use semiconducting materials to convert sunlight directly into electricity. Passive solar systems make use of spontaneous movement of heat by conduction, convection, radiation, or evaporation. (Passive solar systems are not further discussed in this article due to a lack of data.)

Solar energy technologies have important advantages for electric utility and other users. First, the solar resource has no "fuel" cost and is abundantly available. In many areas of the country, the sun shines intensely on a daily basis, providing a nearly unlimited power resource for devices that convert the sun's insolation into useful energy. Second, solar daily output often matches air-conditioning loads and thus provides high value. Third, most solar equipment is modular, allowing supply to be adjusted to demand. Fourth, most PV technologies can easily be operated in remote off-grid areas without transmission and distribution costs. Finally, solar technologies are environmentally clean, with no operating wastes, air pollution, or noise.

Solar energy technologies have some limitations for widespread use. Although costs have declined substantially during the past three decades, they are still very expensive relative to conventional technologies. Also, sunlight is intermittent, resulting in storage and load-matching problems.

7.1 Current Data

Total shipments of solar thermal collectors were 11.2 million square feet in 2001. This represented an

increase of 34% from the 2000 total of 8.4 million square feet. There were 26 companies shipping solar collectors in 2001, the same as in 2000. Import shipments totaled 3.5 million square feet, whereas export shipments totaled 0.8 million square feet. PV cell and module shipments reached 97.7 peak MW in 2001. For the first time in more than 8 years, exports declined, decreasing to 61.4 peak MW. Exports accounted for 63% of total shipments in 2001 compared to 78% in 2000.

7.2 Technology and Applications

Solar energy technologies are classified into two major categories by the type of energy used: solar thermal devices, which use the sun's heat energy, and PV devices, which use the energy inherent in the solar photons and convert it directly into electricity. A solar thermal collector usually consists of a frame, glazing, and an absorber, together with the appropriate insulation. The heat collected by the collector may be used immediately or stored for later use. A PV cell is an electric device consisting of layers of semiconductor materials fabricated to form a junction (adjacent layers of materials with different electric characteristics) and electrical contacts, being capable of converting incident light directly into electricity (direct current). A PV module is two or more cells attached and designed to be used for PV power systems.

7.2.1 Types of Solar Technologies

Data are collected by the EIA on solar-thermal collectors and PV cells and modules. These data provide information on solar energy applications.

7.2.1.1 Solar thermal collector types Solar thermal collectors are classified as low-, medium-, and high-temperature collectors. Low-temperature collectors provide heat up to 110°F through either metallic or nonmetallic absorbers and are used for applications such as swimming pool heating, which has dominated the industry since the rescission of the energy investment tax credits in 1985. Medium-temperature collectors provide heat higher than 110°F (usually 140–180°F) through either glazed flat-plate collectors that use air or liquid as the heat-transfer medium or concentrator collectors that concentrate the heat of incident insulation to greater than one sun. Before the expiration of the energy tax credits at the end of 1985, medium-temperature collector manufacturers dominated the market during the 1970s and early 1980s. Medium-temperature collectors are used for water, space, and process

heating. Evacuated-tube collectors are included in this category. High-temperature collectors are parabolic dish and trough collectors used primarily by utilities and nonutility power producers in the generation of electricity for the grid. High-temperature solar thermal collectors, which operate at temperatures higher than 180°F, are used primarily for R&D projects.

7.2.1.2 Photovoltaic cell and module types PV cells and modules are classified into three categories by product type: crystalline silicon cells and modules (including single-crystal, cast silicon, and ribbon silicon), thin-film cells and modules (made from a number of layers of photosensitive materials such as amorphous silicon), and concentrator cells and modules (in which a lens is used to gather and converge sunlight onto the cells or module surface). Crystalline silicon cells and modules dominate the PV industry.

7.2.2 Uses of Solar Technologies

7.2.2.1 Solar thermal collectors Since 1986, the largest end use for solar thermal collectors has been for heating swimming pools. In 2001, solar swimming pool heaters represented 97% of solar thermal collector shipments (Table V). This application usually involved the use of low-temperature collectors. A common low-temperature pool-heating solar collector is a black plastic or rubber-like sheet with tubing through which water is circulated. The heat of the sun is transferred directly from the black absorbing material to the water circulating through the tubing to supply heat to the pool.

The second largest end use was for domestic hot water systems, which represented 2% of solar collector shipments (Table V). Typical solar water-heating systems feature flat-plate collectors or collectors installed in an internal collector storage or thermosiphon system. Unlike pool-heating systems, domestic solar water-heating systems almost always have a conventional backup (i.e., gas or electric). Medium-temperature collectors also were shipped for space heating and for installation into systems that provide both space and water heating and process heating. High-temperature parabolic dish and trough collectors were shipped for electricity generation and hot water in the past when electricity and natural gas prices were expected to increase very rapidly.

TABLE V

Shipments of Solar Collectors by Market Sector, End Use, and Type, 2000 and 2001 (in Thousands of Square Feet)^a

Type	Low-temperature liquid/air metallic and nonmetallic	Medium-temperature liquid					High-temperature parabolic dish/ trough	2001 total	2000 total
		Air	Integral collector storage/thermosiphon	Flat-plate (pumped)	Evacuated tube	Concentrator			
Market sector									
Residential	9885	3	80	154	1	<i>b</i>	0	10,125	7473
Commercial	987	1	<i>b</i>	22	<i>b</i>	0	1	1012	810
Industrial	12	0	0	5	0	0	0	17	57
Utility	0	0	0	0	0	0	1	1	5
Other	34	0	1	0	0	0	0	35	10
Total	10,919	4	81	181	2	<i>b</i>	2	11,189	8354
End use									
Pool heating	10,782	0	0	16	0	0	0	10,797	7863
Hot water	42	0	81	149	1	<i>b</i>	0	274	367
Space heating	61	4	0	5	<i>b</i>	0	0	70	99
Space cooling	0	0	0	0	0	0	0	0	0
Combined space and water	0	0	0	12	0	0	0	12	2
Process heating	34	0	0	0	0	0	0	34	20
Electricity generation	0	0	0	0	0	0	2	2	3
Other	0	0	0	0	0	0	0	0	0
Total	10,919	4	81	181	2	<i>b</i>	2	11,189	8354

^a Source. Energy Information Administration, Form EIA-63A, "Annual Solar Thermal Collector Manufacturers Survey."^b Less than 500 ft².

TABLE VI

Shipments of Photovoltaic Cells and Modules by Market Sector, End Use, and Type, 2000 and 2001 (in Peak Kilowatts)^a

Sector and end use	Crystalline silicon	Thin-film silicon	Concentrator silicon	Other	2001 total	2000 total
Market						
Industrial	24,754	3135	174	0	28,063	28,808
Residential	28,307	4955	^b	0	33,262	24,814
Commercial	13,440	2270	0	0	15,710	13,692
Transportation	7525	961	0	0	8486	5502
Utility	4799	747	300	0	5846	6298
Government	5375	353	0	0	5728	4417
Other	451	120	0	0	571	4690
Total	84,651	12,541	474	0	97,666	88,221
End use						
Electrical generation						
Grid interactive	22,444	4482	300	0	27,226	21,713
Remote	18,772	2501	174	0	21,447	14,997
Communications	12,974	1769	0	0	14,743	12,269
Consumer goods	3664	395	0	0	4059	2870
Transportation	10,803	1833	0	0	12,636	12,804
Water pumping	6737	708	0	0	7444	5644
Cells/modules to OEM	5778	490	0	0	6268	12,153
Health	2931	272	0	0	3203	2742
Other	549	92	0	0	641	3028
Total	84,651	12,541	474	0	97,666	88,221

^aSource. Energy Information Administration, Form EIA-63B, "Annual Photovoltaic Module/Cell Manufacturers Survey."^bLess than 500 ft².

7.2.2.2 Uses of photovoltaic cells and modules The largest end-use application of PV cells and modules is for electricity generation (combined grid-interactive and remote). Grid-interactive and remote (i.e., stand-alone) power generation includes applications for grid distribution and general remote uses, such as residential power and power for mobile homes. In 2001, electric generation accounted for 50% of total PV shipments (Table VI).

7.3 Niche Markets and Outlook

Solar technologies require either significant cost reductions or a combination of cost reductions and an increase in natural gas prices to become cost competitive in most grid applications. Some obstacles slowing commercialization are technology specific, whereas others are more general. The primary obstacle is that solar technologies cannot currently compete with conventional fossil-fueled technologies in grid-connected applications. On the other hand, some niche applications, such as photovoltaics for transmission and distribution support at the end of a fully loaded distribution line, have value above that

of a new generating plant. Such applications may permit the technology to establish itself and benefit from economies of scale and learning effects.

Acknowledgments

I thank Mark Gielecki, Louise Guey-Lee, and James Holihan for their contributions to this article.

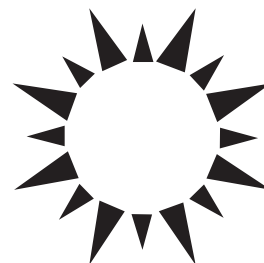
SEE ALSO THE FOLLOWING ARTICLES

Biomass for Renewable Energy and Fuels • Computer Modeling of Renewable Power Systems • Energy Development on Public Land in the United States • National Energy Policy: United States • Public Reaction to Renewable Energy Sources and Systems • Renewable Energy and the City • Renewable Energy in Europe • Renewable Energy in Southern Africa • Renewable Energy Policies and Barriers • Renewable Energy, Taxonomic Overview • Renewable Portfolio Standard

Further Reading

- Energy Information Administration (2001). Renewable energy issues & trends 2000, Report No. DOE/EIA-0628. Energy Information Administration, Washington, DC.
- Energy Information Administration (2002). Renewable energy annual 2001, Report No. DOE/EIA-0603. Energy Information Administration, Washington, DC.
- Energy Information Administration (2003). Annual energy outlook 2003, Report No. DOE/EIA-0383. Energy Information Administration, Washington, DC.
- Federal Energy Regulatory Commission hydropower homepage (2003). <http://www.ferc.gov>.
- Geo-Heat Center (2003). <http://geoheat.oit.edu>.
- U.S. Department of Energy Geothermal Energy Program (2003). <http://www.eere.energy.gov>.
- U.S. Department of Energy Hydropower Program (2003). <http://hydropower.inel.gov>.

Renewable Energy Policies and Barriers



FREDRIC BECK

Renewable Energy Policy Project
Washington, D.C., United States

ERIC MARTINOT

Global Environment Facility
Washington D.C., United States

-
1. Barriers to Renewable Energy
 2. Renewable Energy Promotion Policies
 3. Transport Biofuels Policies
 4. Emissions Reduction Policies
 5. Power-Sector Restructuring Policies
 6. Distributed Generation Policies
 7. Rural Electrification Policies
 8. Summary

Glossary

biofuels Liquid fuels used in transportation that are derived from renewable biomass resources. Not limited to transportation, these fuels can also be used for heating and electricity generation. The main fuels promoted are ethanol, gasohol, biodiesel, bio-oil, and methanol.

capital costs The initial investment costs associated with a renewable energy installation, as distinct from operating and maintenance (all technologies) or fuel costs (in the case of biomass).

distributed generation Generation of electricity at or near the point of end use, often in small quantities, rather than remotely in a large centralized power plants.

independent power producer An organization that produces electric power (i.e., from a wind farm) and sells that power under contract (called a power purchase agreement) to the electric utility or to a third party. If to a third party, the electricity must be transferred over the utility transmission system, usually for a fee.

microcredit Small amounts of credit made available to households or small businesses for investments in high-capital-cost renewable energy supplies or services, usually within a rural development context.

power sector restructuring A political/policy process causing changes in the institutions, ownership, rates and tariffs, power market rules, competition, or financing in the electric power industry (primarily among power generators and electric utilities).

The 1990s saw an explosion of energy policy changes around the globe. Driven by economic, environmental, security, and social concerns, energy regulation has been in great flux. Many of the changes are having a profound influence on renewable energy, both from policies explicitly designed to promote renewable energy and from other policies that indirectly influence incentives and barriers for renewable energy. This article considers six different types of policies that affect renewable energy development, both directly or indirectly: renewable energy promotion policies, transport biofuels policies, emissions reduction policies, electric power restructuring policies, distributed generation policies, and rural electrification policies. Each policy reduces one or more key barriers that impede the development of renewable energy. These barriers are discussed first. In general, most renewable energy policies address cost-related barriers in some manner. Many policies address the requirements for utilities to purchase renewable energy from power producers. Most policies also address the perceived risks of renewable energy in one form or another (i.e., technical, financial, legal). Still others primarily address regulatory and institutional barriers. Some related policies may heighten barriers to renewable energy rather than reduce them.

1. BARRIERS TO RENEWABLE ENERGY

The need for enacting policies to support renewable energy is often attributed to a variety of barriers, or conditions that prevent investments from occurring. Often the result of barriers is to put renewable energy at an economic, regulatory, or institutional disadvantage relative to other forms of energy supply. Barriers include subsidies for conventional forms of energy, high initial capital costs coupled with lack of fuel-price risk assessment, imperfect capital markets, lack of skills or information, poor market acceptance, technology prejudice, financing risks and uncertainties, high transactions costs, and a variety of regulatory and institutional factors. Many of these barriers could be considered market distortions that unfairly discriminate against renewable energy, while others have the effect of increasing the costs of renewable energy relative to the alternatives. Barriers are often quite situation specific in any given region or country; nevertheless, three broad categories of barriers are discussed in this section. Table I summarizes the key renewable energy policies and barriers presented.

1.1 Costs and Pricing

Many argue that renewable energy costs more than other energy sources, resulting in cost-driven decisions and policies that avoid renewable energy. In practice, a variety of factors can distort the comparison. For example, public subsidies may lower the costs of competing fuels. Although it is true that initial capital costs for renewable energy technologies are often higher on a cost-per-unit basis (i.e., \$/kW), it is widely accepted that a true comparison must be made on the basis of total life-cycle costs. Life-cycle costs account for initial capital costs, future fuel costs, future operation and maintenance costs, decommissioning costs, and equipment lifetime. Here lies part of the problem in making comparisons: What are fuel costs going to be in the future? How should future costs be discounted to allow comparison with present costs based on expected interest rates? The uncertainties inherent in these questions affect cost comparisons. Existing analytical tools for calculating and comparing costs can discriminate against renewable energy if they do not account for future uncertainties or make unrealistic assumptions.

Many policies attempt to compensate for cost-related barriers by providing additional subsidies for

renewable energy in the form of tax credits or incentives, by establishing special pricing and power-purchasing rules, and by lowering transaction costs. Despite many calls for reducing subsidies for fossil fuels and nuclear power, in practice this proves politically difficult. Thus, practical policies have tended to focus on increasing subsidies for renewable energy rather than reducing subsidies for fossil fuels and nuclear power.

1.1.1 Subsidies for Competing Fuels

Large public subsidies, both implicit and explicit, are channeled in varying amounts to all forms of energy, which can distort investment cost decisions. Organizations such as the World Bank and International Energy Agency put global annual subsidies for fossil fuels in the range of \$100 billion to \$200 billion, although such figures are very difficult to estimate (for comparison, the world spends some \$1 trillion annually on purchases of fossil fuels). Public subsidies can take many forms: direct budgetary transfers, tax incentives, R&D spending, liability insurance, leases, land rights-of-way, waste disposal, and guarantees to mitigate project financing or fuel price risks. Large subsidies for fossil fuels can significantly lower final energy prices, putting renewable energy at a competitive disadvantage if it does not enjoy equally large subsidies.

1.1.2 High Initial Capital Costs

Even though lower fuel and operating costs may make renewable energy cost competitive on a life-cycle basis, higher initial capital costs can mean that renewable energy provides less installed capacity per initial dollar invested than conventional energy sources. Thus, renewable energy investments generally require higher amounts of financing for the same capacity. Depending on the circumstances, capital markets may demand a premium in lending rates for financing renewable energy projects because more capital is being risked up front than in conventional energy projects. Renewable energy technologies may also face high taxes and import duties. These duties may exacerbate the high first-cost considerations relative to other technologies and fuels.

1.1.3 Difficulty of Fuel Price Risk Assessment

Risks associated with fluctuations in future fossil fuel prices may not be quantitatively considered in decisions about new power generation capacity because these risks are inherently difficult to assess. Historically, future fuel price risk has not been considered an important factor because future fossil

fuel prices have been assumed to be relatively stable or moderately increasing. Thus, risks of severe fluctuations are often ignored. However, with greater geopolitical uncertainties and energy market deregulation has come new awareness about future fuel price risks. Renewable energy technologies avoid fuel costs (with the exception of biomass) and so avoid fuel price risk. However, this benefit, or risk-reduction premium, is often missing from economic comparisons and analytical tools because it is difficult to quantify. Further, for some regulated utilities, fuel costs are factored into regulated power rates so that consumers rather than utilities bear the burden of fuel price risks, and utility investment decisions are made without considering fuel price risk.

1.1.4 Unfavorable Power Pricing Rules

Renewable energy sources feeding into an electric power grid may not receive full credit for the value of their power. Two factors are at work. First, renewable energy generated on distribution networks near final consumers rather than at centralized generation facilities may not require transmission and distribution (i.e., would displace power coming from a transmission line into a node of a distribution network). But utilities may only pay wholesale rates for the power, as if the generation was located far from final consumers and required transmission and distribution. Thus, the locational value of the power is not captured by the producer. Second, renewable energy is often an intermittent source whose output level depends on the resource (i.e., wind and sun) and cannot be entirely controlled. Utilities cannot count on the power at any given time and may lower prices for it. Lower prices take two common forms: (1) a zero price for the capacity value of the generation (utility only pays for the energy value) and (2) an average price paid at peak times (when power is more valuable), which is lower than the value of the power to the utility—even though the renewable energy output may directly correspond with peak demand times and thus should be valued at peak prices.

1.1.5 Transaction Costs

Renewable energy projects are typically smaller than conventional energy projects. Projects may require additional information not readily available, or they may require additional time or attention to financing or permitting because of unfamiliarity with the technologies or uncertainties over performance. For these reasons, the transaction costs of renewable

energy projects—including resource assessment, siting, permitting, planning, developing project proposals, assembling financing packages, and negotiating power-purchase contracts with utilities—may be much larger on a per kilowatt (kW) capacity basis than for conventional power plants. Higher transaction costs are not necessarily an economic distortion in the same way as some other barriers, but simply make renewables more expensive. However, in practice some transaction costs may be unnecessarily high, for example, overly burdensome utility interconnection requirements and high utility fees for engineering reviews and inspection.

1.1.6 Environmental Externalities

The environmental impacts of fossil fuels often result in real costs to society, in terms of human health (i.e., loss of work days, health care costs), infrastructure decay (i.e., from acid rain), declines in forests and fisheries, and perhaps ultimately, the costs associated with climate change. Dollar costs of environmental externalities are difficult to evaluate and depend on assumptions that can be subject to wide interpretation and discretion. Although environmental impacts and associated dollar costs are often included in economic comparisons between renewable and conventional energy, investors rarely include such environmental costs in the bottom line used to make decisions.

1.2 Legal and Regulatory

1.2.1 Lack of Legal Framework for Independent Power Producers

In many countries, power utilities still control a monopoly on electricity production and distribution. In these circumstances, in the absence of a legal framework, independent power producers may not be able to invest in renewable energy facilities and sell power to the utility or to third parties under so-called power purchase agreements. Or utilities may negotiate power purchase agreements on an individual ad hoc basis, making it difficult for project developers to plan and finance projects on the basis of known and consistent rules.

1.2.2 Restrictions on Siting and Construction

Wind turbines, rooftop solar hot-water heaters, photovoltaic installations, and biomass combustion facilities may all face building restrictions based on height, aesthetics, noise, or safety, particularly in urban areas. Wind turbines have faced specific

TABLE I

Summary of Renewable Energy Policies and Barriers

Policies	Description	Key barriers addressed
Renewable Energy Promotion Policies		
Price-setting and quantity-forcing policies	Mandates prices to be paid for renewable energy or requires a fixed amount or share of generation to be renewable	High costs, unfavorable power pricing rules, perceived risks
Cost-reduction policies	Reduces investment costs through subsidies, rebates, tax relief, loans and grants	High costs, perceived risks
Public investments and market facilitation activities	Provides public funds for direct investments or for guarantees, information, training, etc. to facilitate investments	Transaction costs, perceived risks, lack of access to credit, information, and skills
Power grid access policies	Gives renewable energy equal or favorable treatment for access to power grids and transmission systems	Independent power producer frameworks, transmission access, interconnection requirements
Transport Biofuels Policies		
Biofuels mandates	Mandates specific shares of transport fuel consumption from biofuels	Lack of fuel production or delivery infrastructure
Biofuel tax policies	Provides tax relief for biofuels	High costs
Emissions Reduction Policies		
Renewable energy set-asides	Allocates, or sets aside, a percentage of mandated environmental emissions reductions to be met by renewable energy	Environmental externalities
Emissions cap and trade policies	Allows renewables to receive monetary credit for local pollutant emissions reductions	Environmental externalities
Greenhouse gas mitigation policies	Allows renewables to receive monetary credit for greenhouse-gas emissions reductions	Environmental externalities
Power-Sector Restructuring Policies		
Competitive wholesale power markets	Allows competition in supplying wholesale generation to the utility network and eliminates wholesale pricing restrictions.	May heighten barriers of high costs, lack of fuel price risk assessment, unfavorable power pricing rules
Self-generation by end users	Allows end users to generate their own electricity and either sell surplus power back to the grid or partly offset purchased power	May reduce barrier of interconnection requirements, but heighten barriers of high costs, lack of fuel price risk assessment
Privatization or commercialization of utilities	Changes government-owned and operated utilities into private or commercial entities	May reduce barrier of subsidies, but heighten barriers of high capital costs and perceived risks
Unbundling of generation, transmission and distribution	Eliminates monopolies so that separate entities provide generation, transmission, and distribution	May provide greater incentives to self-generate, including with renewable energy
Competitive retail power markets	Provides competition at the retail level for power sales, including "green power" sales	May reduce barriers of high costs, lack of information, transaction costs
Distributed Generation Policies		
Net metering	Values renewable energy production at the point of end use and allows utility networks to provide "energy storage" for small users	Unfavorable power pricing rules
Real-time pricing	Values renewable energy production at the actual cost of avoided fossil fuel generation at any given time of the day	Unfavorable power pricing rules
Capacity credit	Provides credit for the value of standing renewable energy capacity, not just energy production	Unfavorable power pricing rules

continues

Table 1 continued

Policies	Description	Key barriers addressed
Interconnection regulations	Creates consistent and transparent rules, norms, and standards for interconnection	Interconnection requirements, transaction costs
Rural Electrification Policies		
Rural electrification policy and energy service concessions	Makes renewable energy part of rural electrification policy and planning and creates regulated businesses to serve rural customers	Subsidies for competing fuels, lack of skills and information, high costs, lack of access to credit
Rural business development and microcredit	Supports private entrepreneurs to provide renewable energy products and services to end users and offer consumer credit for purchases	Lack of skills, lack of access to credit
Comparative line extension analyses	Analyzes the relative costs of renewable energy with conventional fuels and power delivery	Subsidies for competing fuels, lack of information

environmental concerns related to siting along migratory bird paths and coastal areas. Urban planning departments or building inspectors may be unfamiliar with renewable energy technologies and may not have established procedures for dealing with siting and permitting. Competition for land use with agricultural, recreational, scenic, or development interests can also occur.

1.2.3 Transmission Access

Utilities may not allow favorable transmission access to renewable energy producers or may charge high prices for transmission access. Transmission access is necessary because some renewable energy resources like windy sites and biomass fuels may be located far from population centers. Transmission or distribution access is also necessary for direct third-party sales between the renewable energy producer and a final consumer. New transmission access to remote renewable energy sites may be blocked by transmission-access rulings or right-of-way disputes.

1.2.4 Utility Interconnection Requirements

Individual home or commercial systems connected to utility grids can face burdensome, inconsistent, or unclear utility interconnection requirements. Lack of uniform requirements can add to transaction costs. Safety and power-quality risk from nonutility generation is a legitimate concern of utilities, but a utility may tend to set interconnection requirements that go beyond what is necessary or practical for small producers in the absence of any incentive to set more reasonable but still technically sound requirements. In turn, the transaction costs of hiring legal and technical experts to understand and comply with interconnection requirements may be significant. Policies that create sound and uniform interconnection standards can reduce interconnection hurdles and costs.

1.2.5 Liability Insurance Requirements

Small power generators (particularly home photovoltaic systems feeding into the utility grid under so-called net metering provisions) may face excessive requirements for liability insurance. The phenomenon of islanding, which occurs when a self-generator continues to feed power into the grid when power flow from the central utility source has been interrupted, can result in serious injury or death to utility repair crews. Although proper equipment standards can prevent islanding, liability is still an issue. Several U.S. states have prohibited utilities from requiring additional insurance beyond normal homeowner liability coverage as part of net metering statutes.

1.3 Market Performance

1.3.1 Lack of Access to Credit

Consumers or project developers may lack access to credit to purchase or invest in renewable energy because of lack of collateral, poor creditworthiness, or distorted capital markets. In rural areas, microcredit lending for household-scale renewable energy systems may not exist. Available loan terms may be too short relative to the equipment or investment lifetime. In some countries, power project developers have difficulty obtaining bank financing because of uncertainty as to whether utilities will continue to honor long-term power purchase agreements to buy the power.

1.3.2 Perceived Technology Performance Uncertainty and Risk

Proven, cost-effective technologies may still be perceived as risky if there is little experience with them in a new application or region. The lack of visible installations and familiarity with renewable

energy technologies can lead to perceptions of greater technical risk than for conventional energy sources. These perceptions may increase required rates of return, result in less capital availability, or place more stringent requirements on technology selection and resource assessment. “Lack of utility acceptance” is a phrase used to describe the historical biases and prejudices on the part of traditional electric power utilities. Utilities may be hesitant to develop, acquire, and maintain unfamiliar technologies, or give them proper attention in planning frameworks. Finally, prejudice may exist because of poor past performance that is out of step with current performance norms.

1.3.3 Lack of Technical or Commercial Skills and Information

Markets function best when everyone has low-cost access to good information and the requisite skills. But in specific markets, skilled personnel who can install, operate, and maintain renewable energy technologies may not exist in large numbers. Project developers may lack sufficient technical, financial, and business development skills. Consumers, managers, engineers, architects, lenders, or planners may lack information about renewable energy technology characteristics, economic and financial costs and benefits, geographical resources, operating experience, maintenance requirements, sources of finance, and installation services. The lack of skills and information may increase perceived uncertainties and block decisions.

2. RENEWABLE ENERGY PROMOTION POLICIES

Policies whose specific goal is to promote renewable energy fall into three main categories: (1) price-setting and quantity-forcing policies, which mandate prices or quantities; (2) investment cost reduction policies, which provide incentives in the form of lower investment costs; and (3) public investments and market facilitation activities, which offer a wide range of public policies that reduce market barriers and facilitate or accelerate renewable energy markets. Historically, governments have enacted these policies in a rather ad hoc manner. More recently, national renewable energy targets (also referred to as goals) have emerged as a political context for promoting specific combinations of policies from all three categories. Such targets focus on the

aggregate energy production of an entire country or group of countries. Targets may specify total primary energy from renewables or minimum renewable energy shares of electricity generation.

Several countries have adopted or are proposing national renewable energy targets. The European Union collectively has adopted a target of 22% of total electricity generation from renewables by 2010, with individual member states having individual targets above or below that amount. Japan has adopted a target of 3% of total primary energy by 2010. Recent legislative proposals in the United States would require 10% of electricity generation from renewables by 2020. China and India are the first developing countries to propose renewable energy targets. India has proposed that by 2012, 10% of annual additions to power generation would be from renewable energy; China has a similar goal of 5% by 2010. Other countries with existing or proposed targets are Australia, Brazil, Malaysia, and Thailand. In addition, countries from around the world placed increased attention on renewable energy targets at the United Nations World Summit for Sustainable Development in 2002.

2.1 Price-Setting and Quantity-Forcing Policies

Price-setting policies reduce cost- and pricing-related barriers by establishing favorable pricing regimes for renewable energy relative to other sources of power generation. The quantity of investment obtained under such regimes is unspecified, but prices are known in advance. Quantity-forcing policies do the opposite; they mandate a certain percentage or absolute quantity of generation to be supplied from renewable energy, at unspecified prices. Often price-setting or quantity-forcing policies occur in parallel with other policies, such as investment cost-reduction policies.

The two main price-setting policies seen to date are the PURPA legislation in the United States and electricity feed-in laws in Europe. The two main quantity-forcing policies seen to date are competitively bid renewable-resource obligations and renewable portfolio standards.

2.1.1 U.S. Public Utility Regulatory Policies Act (PURPA)

PURPA was enacted in 1978 in part to encourage electric power production by small power producers using renewable resources to reduce U.S. dependence

on foreign oil. The policy required utilities to purchase power from small renewable generators and cogenerators, known as qualifying facilities, through long-term (10-year) contracts at prices approximating the avoided costs of the utilities. These avoided costs represented the marginal costs to the utilities of building new generation facilities, which could be avoided by purchasing power from the qualifying facilities instead. Avoided cost calculations typically assumed an aggressive schedule of escalating future energy prices, making contract prices to qualifying facilities quite attractive.

For example, standard offer contracts under PURPA in California in the 1980s set the avoided cost of generation in the range 6 to 10 cents/kWh—very favorable rates, which initially spurred many investments by renewable energy producers. However, by the 1990s energy prices had not risen as originally expected and a large number of natural-gas-fired independent generation came online in California. Power surpluses emerged, wholesale power prices declined, and declining standard offer rates led to reduced competitiveness of renewable energy and a significant slowdown in construction of new capacity.

2.1.2 *Electricity Feed-in Laws*

The electricity feed-in laws in Germany, and similar policies in other European countries in the 1990s, set a fixed price for utility purchases of renewable energy. For example, in Germany starting in 1991, renewable energy producers could sell their power to utilities at 90% of the retail market price. The utilities were obligated to purchase the power. The German feed-in law led to a rapid increase in installed capacity and development of commercial renewable energy markets. Wind power purchase prices were highly favorable, amounting to about DM 0.17/kWh (US 10 cents/kWh), and applied over the entire life of the plant. Total wind power installed went from near zero in the early 1990s to over 8500 MW by 2001, making Germany the global leader in renewable energy investment.

Partly because retail electricity prices declined with increasing competition due to electricity deregulation, which made producers and financiers wary of new investments, a new German Renewable Energy Law of 2000 changed electricity feed-in pricing. Pricing became based on fixed norms unique to each technology, which in turn were based on estimates of power production costs and expectations of declines in those costs over time. For example, wind power prices remained at the previous

level of DM 0.17/kWh for plants commissioned in 2001, but only for the first 5 years of operation, after which prices paid declined. Solar PV prices were set initially at DM 0.99/kWh. All prices had built-in declines over time (i.e., 1.5% annual decreases in starting tariffs paid for wind power plants commissioned in subsequent years). This provision addressed one of the historical criticisms of feed-in approaches, which was that they did not encourage technology cost reductions or innovation. The new law's provisions for regular adjustments to prices addressed technological and market developments. The law also distributed the costs of the policy (i.e., the additional costs of wind power over conventional power) among all utility customers in the country. This issue of burden sharing had become a significant political factor in Germany by 2000 because the old law placed a disproportionate burden on utility customers in specific regions where wind power development was heaviest.

Other countries in Europe with renewable electricity feed-in laws include Denmark, France, Greece, Italy, Portugal, Spain, and Sweden. A combination of feed-in tariffs, production subsidies of DK 0.10/kWh, and a strong domestic market helped the Danish wind industry maintain a 50% market share of global wind turbine production for a number of years.

2.1.3 *Competitively-Bid Renewable-Resource Obligations*

The United Kingdom tried competitive bidding for renewable-energy-resource obligations during the 1990s under its Non-Fossil-Fuel Obligation (NFFO) policy. Under the NFFO, power producers bid on providing a fixed quantity of renewable power, with the lowest-price bidder winning the contract. With each successive bidding round (there were four total), bidders reduced prices relative to the last round. For example, wind power contract prices declined from 10 p/kWh in 1990 under NFFO-1 to 4.5 p/kWh in 1997 under NFFO-4. One of the lessons some have drawn from the UK experience is that competitively determined subsidies can lead to rapidly declining prices for renewable energy. However, there has also been criticism that the NFFO process encouraged competing projects to bid below cost in order to capture contracts, with the result that successful bidders were unable to meet the terms of the bid or ended up insolvent. This criticism proved valid in practice; contracts awarded to low bidders did not always translate into projects on the ground. The UK abandoned the NFFO approach after the fourth

round of bidding in 1997. Other countries with similar competitively bid renewable resource mechanisms have included Ireland (under the Alternative Energy Requirement program), France (under the EOLE program), and Australia (under the Renewable Energy Commercialisation Program program).

2.1.4 Renewable Energy Portfolio Standards (RPS)

An RPS requires that a minimum percentage of generation sold or capacity installed be provided by renewable energy. Obligated utilities are required to ensure that the target is met, either through their own generation, power purchases from other producers, or direct sales from third parties to the utility's customers. Typically, RPS obligations are placed on the final retailers of power, who must purchase either a portion of renewable power or the equivalent amount of green certificates (see the next section). Two types of standards have emerged: capacity-based standards set a fixed amount of capacity by a given date, while generation-based standards mandate a given percentage of electricity generation that must come from renewable energy.

In the United States, many RPS policies have occurred as part of utility restructuring legislation. These are typically generation-based standards with phased implementation to allow utilities to reach incrementally increasing targets over a number of years. At least 12 U.S. states have enacted an RPS, ranging from 1 to 30% of electricity generation. However, the amount of new and additional generation expected from these standards varies widely depending on existing renewable energy capacity. For example, Maine historically generates over 40% of its power from renewable resources, so its 30% standard is unlikely to result in any new renewable generation. In contrast, California's requirement to increase renewable sales from 10.5% in 2001 to 20% by 2017 will likely result in a significant amount of new in-state renewable energy generation. Texas implemented an RPS in 2000 requiring 2000 MW of new renewable capacity by 2008. Partly due to federal production tax credits, Texas has been substantially ahead of schedule, with half of the targeted capacity in place by 2002.

In Europe, the Netherlands has been a leader among RPS initiatives. Dutch utilities have adopted an RPS voluntarily, based on targets of 5% of electricity generation by 2010, increasing to 17% by 2020. Other countries with regulatory requirements for utilities or electricity retailers to purchase a percentage of renewable power include Australia, Brazil, Belgium, Denmark, France, Japan, Spain,

Sweden, and the United Kingdom. The UK's Renewables Obligation on suppliers will rise in annual steps from 3% in 2003 to 10% in 2010. In Denmark, legislation obliges end users to purchase 20% of their electricity from renewable sources by 2003.

In Brazil, a policy enacted during the electricity crisis of 2001 requires national utilities to purchase more than 3000 MW of renewable energy capacity by 2016. Purchase prices are set by the government at 80% of the national average electricity retail price. Thus, in contrast to most national policies elsewhere, Brazil's policy is effectively both price setting and quantity forcing.

2.1.5 Renewable Energy (Green) Certificates

Renewable energy (green) certificates are emerging as a way for utilities and customers to trade renewable energy production or consumption credits in order to meet obligations under RPS and similar policies. Standardized certificates provide evidence of renewable energy production and are coupled with institutions and rules for trading that separate renewable attributes from the associated physical energy. This enables a paper market for renewable energy to be created independent of actual electricity sales and flows. Green certificate markets are emerging in several countries, allowing producers or purchasers of renewable energy who earn green certificates to sell those certificates to those who need to meet obligations but have not generated or purchased the renewable power themselves. Those without obligations but who wish to voluntarily support green power for philosophical or public relations reasons may also purchase certificates.

Public and private institutions are emerging that keep track of renewable energy generation, assign certificates to generators, and register trades and sales of certificates. Green certificate trading is gaining ground in the UK, Belgium, Denmark, Australia, and the United States. Europe embarked on a test phase of an EU-wide renewable energy certificate trading system during 2001 and 2002. This Renewable Energy Certificate System (RECS) was established as a formal association in December 2002 with over 100 members in 14 different countries. Certificates for over 22,500 GWh were issued through 2003.

2.2 Cost-Reduction Policies

A number of policies are designed to provide incentives for voluntary investments in renewable energy by reducing the costs of such investments.

These policies can be characterized as falling in five broad categories: policies that (1) reduce capital costs up front (via subsidies and rebates), (2) reduce capital costs after purchase (via tax relief), (3) offset costs through a stream of payments based on power production (via production tax credits), (4) provide concessionary loans and other financial assistance, and (5) reduce capital and installation costs through economies of bulk procurement.

2.2.1 Subsidies and Rebates

Reduction in the initial capital outlay by consumers for renewable energy systems is accomplished through direct subsidies, or rebates. These subsidies are used to buy down the initial capital cost of the system so that the consumer sees a lower price. In the United States, at least nineteen states offer rebate programs at the state, local, or utility level to promote the installation of renewable energy equipment. The majority of the programs are available from state agencies and municipally owned utilities and support solar water heating or photovoltaic systems, though some include geothermal heat pumps, small wind generators, passive solar, biomass, and fuel cells. Homes and businesses are usually eligible, although some programs target industry and public institutions as well. In some cases, rebate programs are combined with low- or no-interest loans.

Sustained efforts to increase the use of renewables have been made via coordinated, multiyear, multi-policy initiatives. For example, Japan, Germany, and the United States subsidize capital costs of solar PV as part of their market transformation program.

- Japan's Sunshine Program provides capital subsidies and net metering for rooftop PV systems. From 1994 to 2000, the government invested 86 billion yen (\$725 million USD), resulting in 58,000 system installations and over 220 MW of PV capacity. Subsidies began at 900,000 yen/peak-kW (US\$5/peak-watt) in 1994, and were gradually reduced to 120,000 yen/peak-kW (US\$1/peak-watt) in 2001 as PV prices fell.

- Germany began a 1000 solar roofs program in 1991 that offered subsidies for individual household purchases of solar PV of up to 60% of capital system costs. The program was expanded in 1999 to 100,000 roofs over 5 years, providing 10-year low-interest loans to households and businesses. As a result of favorable feed-in tariffs and low-interest loans, the program was expected to provide 300 peak-MW of PV capacity.

- The United States launched a million solar roofs initiative in 1997 to install solar PV systems and solar thermal systems on 1 million buildings by 2010. The program includes long-term, low-interest customer financing, government procurement for federal buildings, commercialization programs, and production incentives. Individual states also have capital subsidy programs for PV, with the California Energy Commission offering rebates of up to \$4.50/peak-watt or 50% off the system purchase price, New York and New Jersey offering up to \$5/peak-watt subsidies, and New York rebating as much as 70% of the cost of eligible equipment as of 2002.

Subsidy programs also exist for wind power, such as Denmark's DK 0.10/kWh (US 1.5 cents/kWh) production subsidy paid to utilities. Among developing countries, Thailand provided subsidies for small renewable energy power producers starting in 2000, soliciting bids for 300 MW of small renewable power and providing production subsidies above standard power purchase rates for at least the first 5 years of operation of each facility.

2.2.2 Tax Relief

Tax relief policies to promote renewable energy have been employed in the United States, Europe, Japan, and India. Tax relief has been especially popular in the United States, where a host of federal and state tax policies address energy production, property investments, accelerated depreciation, and renewable fuels. State policies vary widely in scope and implementation. At least 17 states have personal tax incentives, 21 states have corporate tax incentives, 16 states have sales tax incentives, and 24 states have property tax incentives.

2.2.2.1 Investment Tax Credits

Investment tax credits for renewable energy have been offered for businesses and residences. In the United States, businesses receive a 10% tax credit for purchases of solar and geothermal renewable energy property, subject to certain limitations. Some U.S. states have investment tax credits of up to 35%.

2.2.2.2 Accelerated Depreciation

Accelerated depreciation allows renewable energy investors to receive the tax benefits sooner than under standard depreciation rules. The effect of accelerated depreciation is similar to that of investment tax credits. In the United States, businesses can recover investments in solar, wind, and geothermal property by depreciating them over a period of 5 years, rather than the

15- to 20-year depreciation lives of conventional power investments.

India's accelerated depreciation policy allowed 100% depreciation in the first year of operation, helping spur the largest wind power industry among developing countries. However, this policy led to large investments without sufficient regard to long-term operating performance and maintenance, resulting in capacity factors lower than for wind power installations elsewhere. This led many to conclude that production-based incentives are preferable to investment tax credits and accelerated depreciation, although Germany's investment tax credits accompanied by wind turbine technical standards and certification requirements avoided the problems found in India.

2.2.2.3 Production Tax Credits A production tax credit provides the investor or owner of qualifying property with an annual tax credit based on the amount of electricity generated by that facility. By rewarding production, these tax credits encourage improved operating performance. A production tax credit in Denmark provides DK 0.10/kWh (US 1.5 cents/kWh) for wind power, but few other countries have adopted similar credits. In the United States the Renewable Electricity Production Credit (PTC) provides a per kWh tax credit for electricity generated by qualified wind, closed-loop biomass, or poultry waste resources. Federal tax credits of 1.5 cents/kWh (adjusted annually for inflation) are provided for the first ten years of operation for all qualifying plants that entered service from 1992 through mid-1999, later extended to 2001 and then to 2003.

At least five U.S. states have state or local production incentives for distributed electrical generation, renewable fuels, or both. These policies are similar to the federal PTC, with specific limits on technologies, dates-in-service, and maximum payout per provider and per year. Funds to support the incentives are obtained from a mixture of sources, including general funds, public benefit or environmental funds, and green electricity sales (so-called green tags).

2.2.2.4 Property Tax Incentives At least 24 U.S. states have property tax incentives for renewable energy. These incentives are implemented on many scales—state, county, city, town, and municipality. These are generally implemented in one of three ways: (1) renewable energy property is partially or

fully excluded from property tax assessment, (2) renewable energy property value is capped at the value of an equivalent conventional energy system providing the same service, and (3) tax credits are awarded to offset property taxes.

2.2.2.5 Personal Income Tax Incentives Credits against personal state income taxes are available in many U.S. states for purchase of or conversion to eligible renewable energy systems and renewable fuels. In some cases, taxpayers can deduct the interest paid on loans for renewable energy equipment.

2.2.2.6 Sales Tax Incentives At least 16 U.S. states have policies that provide retail sales tax exemptions for eligible renewable energy systems and renewable fuels. Most exempt 100% of the sales tax for capital expenses and provide specific cents-per-gallon exemptions for renewable fuels. Some policies specify maximum or minimum sizes for eligible systems.

2.2.2.7 Pollution Tax Exemptions The Netherlands is an example where so-called green power is exempt from a new and rising fossil-fuel tax on electricity generation that is paid by end users. Starting in 2001, that fossil-fuel tax rose to the equivalent of US 5 cents/kWh, providing a large tax incentive for Dutch consumers.

2.2.2.8 Other Tax Policies A variety of other tax policies exist, such as income tax exemptions on income from renewable power production, excise duty and sales tax exemptions on equipment purchased, and reduced or zero import tax duties on assembled renewable energy equipment or on components. India, for example, has allowed 5-year tax exemptions on income from sales of wind power.

2.2.3 Grants

Many countries have offered grants for renewable energy purchases. For example, beginning in 1979, Denmark provided rebates of up to 30% of capital costs for wind and other renewable energy technologies. These rebates declined over time. In the United States, county and state governments and utilities provide grants for renewable energy ranging in size from hundreds to millions of dollars.

2.2.4 Loans

Loan programs offer financing for the purchase of renewable energy equipment. Loans can be market rate, low interest (below-market rate), or forgivable.

In many U.S. states, loans are available to virtually all sectors—residential, commercial, industrial, transportation, public, and nonprofit. Repayment schedules vary, with terms of up to 10 years common. Interest rates for renewable energy investments can often be 1% or more higher than those for conventional power projects because of the higher perceived risks involved, so government-subsidized loans that offer below-market interest rates are also common.

Renewable energy loans can take many forms. Residential loans may range from \$500 to \$10,000 or more, while commercial and industrial loans may extend to the millions. Funding comes from a variety of sources, including municipal bonds, system benefit funds, revolving funds, and utility penalty or overcharge funds. Financing may be for a fraction to 100% of a project. Some loan programs have minimum or maximum limits, while others are open ended. Loan terms range from 3 years to the life of a project. Some loans are contractor driven and may include service contracts in the loan amount. Sometimes grants and loans are combined; for example, Iowa provides a 20% forgivable loan combined with an 80% loan at prime rate for renewable fuels projects.

In some developing countries, notably India, China, and Sri Lanka, multilateral loans by lenders such as the World Bank have provided financing for renewable energy, usually in conjunction with commercial lending. One of the most prominent examples is the India Renewable Energy Development Agency (IREDA), which was formed in 1987 to provide assistance in obtaining international multilateral agency loans and in helping private power investors obtain commercial loans. By 2001, IREDA had disbursed the equivalent of more than US\$400 million in loans for renewable energy projects in India, resulting in more than 1600 MW of renewable power generation.

2.3 Public Investments and Market Facilitation Activities

2.3.1 Public Benefit Funds

In the United States, public funds for renewable energy development are raised through a System Benefits Charge (SBC), which is a per-kWh levy on electric power consumption. Some analysts suggest that state clean energy funds seem to be one of the more effective policies in promoting renewable energy development to result from electricity restruc-

turing. It is estimated that 14 U.S. states will collect \$3.5 billion through 2011 in system benefits charges. Similar levies exist in some European countries for fossil-fuel-based generation. In general, the funds serve a variety of purposes, such as paying for the difference between the cost of renewables and traditional generating facilities, reducing the cost of loans for renewable facilities, providing energy efficiency services, funding public education on energy-related issues, providing low-income energy assistance, and supporting research and development.

2.3.2 Infrastructure Policies

Market facilitation supports market institutions, participants, and rules to encourage renewable energy technology deployment. A variety of policies are used to build and maintain this market infrastructure, including policies for design standards, accelerated siting and permitting, equipment standards, and contractor education and licensing. Additionally, policies to induce renewable technology manufacturers to site locally and direct sales of renewable systems to customers at concessionary rates facilitate market development.

2.3.2.1 Construction and Design Policies Construction and design standards include building-code standards for PV installations, design standards evaluated on life-cycle cost basis, and performance requirements. Policy examples include Tucson, Arizona, which requires that commercial facilities achieve a 50% reduction in energy usage over 1995 Model Energy Code, and Florida, which requires that all new educational facilities include passive solar design.

2.3.2.2 Site Prospecting, Review and Permitting Federal and state programs reduce barriers to renewable energy development through resource, transmission, zoning, and permitting assessments. This particularly helped early promotion of wind energy projects in California. On a national scale the Utility Wind Resource Assessment Program funds a number of supporting activities, including up to 50% of the cost of wind resource assessments. India also has a large wind assessment program, with more than 600 stations in 25 states providing information to project developers on the best sites for development.

2.3.2.3 Equipment Standards and Contractor Certification A variety of equipment-related standards

and certification measures have been applied to ensure uniform quality of equipment and installation, increasing the likelihood of positive returns from renewable energy installations. Contractor licensing requirements ensure that contractors have the necessary experience and knowledge to properly install systems. Equipment certifications ensure that equipment meets certain minimum standards of performance or safety.

2.3.2.4 Industrial Recruitment Industrial recruitment policies use financial incentives such as tax credits, grants, and government procurement commitments to attract renewable energy equipment manufacturers to locate in a particular area. These incentives are designed to create local jobs, strengthening the local economy and tax base, and improving the economics of local renewable development initiatives.

2.3.2.5 Direct Equipment Sales These programs allow the consumer to buy or lease renewable energy systems directly from electric provider at below-retail rates. Some programs provide a capital buydown. Examples include Arizona, which provides a buydown of \$2/peak-watt for PV, and California's Sacramento Municipal Utility District, which offers a 50% buydown plus 10-year financed loans and net metering.

2.3.3 Government Procurement Government procurement policies aim to promote sustained and orderly commercial development of renewable energy. Governmental purchase agreements can reduce uncertainty and spur market development through long-term contracts, preapproved purchasing agreements, and volume purchases. Government purchases of renewable energy technologies in early market stages can help overcome institutional barriers to commercialization, encourage the development of appropriate infrastructure, and provide a market path for technologies that require integrated technical, infrastructure, and regulatory changes.

2.3.4 Customer Education and Mandated Generation Disclosure Information U.S. restructuring and deregulation policies mandate that information be provided to customers about choice of electricity providers and characteristics of electricity being provided (such as emissions and fuel types). In many states, general education to raise customer awareness about renewable energy and the environmental impacts of energy generation is required, typically via websites and printed materials.

2.3.5 Solar and Wind Access Laws

Renewable access laws address access, easements, and covenants. Access laws provide a property owner the right to continued access to a renewable resource. Easements provide a privilege to have continued access to wind or sunlight, even though development or features of another person's property could reduce that access. Easements are often voluntary contracts and may be transferred with the property title. Covenant laws prohibit neighborhood covenants from explicitly restricting the installation or use of renewable equipment. Policy mechanisms include access ordinances, development guidelines addressing street orientation, zoning ordinances with building height restrictions, and renewable permits.

3. TRANSPORT BIOFUELS POLICIES

Biofuels mandates and tax policies in Brazil, the United States, and Europe have supported accelerating development of biofuels. Biofuels mandates require that a certain percentage of all liquid transport fuels be derived from renewable resources. Tax policies may provide tax credits or exemptions for production or purchase of biofuels.

Brazil has been quite successful with biofuels mandates under its ProAlcool program, which has promoted the use ethanol for transportation fuel since the 1980s. In addition to a variety of economic incentives and subsidies, Brazil has mandated that ethanol be blended with all gasoline sold in the country. Brazil has also required that all gas stations sell pure ethanol. This last requirement made it commercially viable for the automotive industry to produce ethanol-only (neat ethanol) cars. However, the share of ethanol-only cars purchased annually, after rising to 95% by 1985, subsequently declined for a number of reasons. Most cars now use the "gasohol" ethanol/gasoline blend, and more than 60% of Brazil's sugar cane production goes to produce 18 billion gallons of ethanol each year, representing 90% of global ethanol production. In 2000, over 40% of automobile fuel consumption in Brazil was ethanol.

The United States, the world's second largest ethanol producer after Brazil, has a number of biofuels tax policies and mandates. The Energy Security Act of 1979 created a federal ethanol tax credit of up to 60 cents per gallon for businesses that sell or use alcohol as a fuel. Gasoline refiners and distributors may also receive an excise tax exemption

of up to 5 cents/gallon for blending their fuels with ethanol. State-level ethanol policies also exist, whose origins in the 1980s can be traced back to initiatives by Iowa to use its corn crop for energy. Several policies in Iowa were established to encourage ethanol consumption, including a mandate for government vehicles to use ethanol-blended fuel, and a 1-cent-per-gallon fuel sales tax exemption for ethanol-blended fuels. In 1998, both the federal government and the state of Iowa extended their ethanol tax exemptions until the year 2007. In part due to ethanol incentives, more than 60 ethanol production facilities have become operational in the United States since 1976, with a production capacity of more than 2.4 billion gallons per year. A recent U.S. legislative proposal, called a renewable fuels standard, would triple biofuel use within 10 years.

In Europe, Germany is the largest user of biodiesel and provides tax incentives for 100% pure biodiesel. These incentives have had a large effect; German consumption of biodiesel went from 200 million gallons in 1991 to 750 million gallons in 2002. Other EU members provide tax incentives for 2 to 5% biodiesel blends. Germany, Austria, and Sweden use 100% pure biodiesel in specially adapted vehicles, and biodiesel is mandated in environmentally sensitive areas in Germany and Austria. Other countries producing biodiesel are Belgium, France, and Italy, the later two also providing tax incentives. The European Commission has recently proposed a biofuel directive with targets or mandates for biofuels up to 6% of all transportation fuel sold.

Among other developing countries, Thailand is considering tax policies for biofuels, including excise tax exemptions for ethanol and income tax waivers for investments in biofuels facilities. Malaysia and Indonesia also utilize biodiesel.

4. EMISSIONS REDUCTION POLICIES

Policies to reduce power plant emissions, including NO_x , SO_x , and CO_2 , have the potential to affect renewable energy development. Many emissions-reduction policies create allowances for certain emissions (representing the right to emit a certain amount of that pollutant). Credits available to renewable energy generation can offset these allowed emissions. Such credits have market value and are often traded to allow electricity generators to comply with emission regulations at least cost. Three innovative examples of U.S. air quality standards, acid rain

prevention programs, and state-level greenhouse-gas reduction initiatives illustrate the potential of emissions reductions policies to affect renewable energy.

4.1 Renewable Energy Set-Asides

To meet National Ambient Air Quality Standards, the U.S. Environmental Protection Agency requires 22 U.S. states and the District of Columbia to reduce NO_x emissions significantly by 2007. States can meet emission reduction targets through actual emission reductions or purchase of emission reduction credits from other states participating in a region-wide NO_x trading program. States can allocate, or set-aside, a percentage of the total state NO_x allowances to energy efficiency and renewable energy. Eligible renewable energy producers receive these set-aside allowances and can sell them to fossil-fuel-based electricity generators to enable those generators to stay within their NO_x allocation. The additional revenue from sales of these set-aside allowances can potentially provide stimulus for renewable energy development, although to date few states have implemented renewable energy set-asides.

4.2 Emissions Cap-and-Trade Policies

Under the 1990 Clean Air Act the United States instituted a cap-and-trade mechanism to reduce SO_2 emissions and facilitate least-cost compliance. Under the Acid Rain subprogram, 300,000 SO_2 emissions allowances (rights to emit SO_2) were set aside for utilities that employed renewable energy or energy-efficiency measures. Allowances were to be earned from 1992 through 1999, allocated at a rate of one allowance (one ton of SO_2 avoided) per 500 MWh of generation produced by renewable energy or avoided through increased energy efficiency. This program was not particularly effective, as only one-tenth of the 300,000 allowances were allocated to energy efficiency or renewable energy. Analysis suggests that allowance prices set by the market were not high enough to justify renewable energy investments. In addition, the program restricted participation to utilities, excluding other power generators, which also contributed to under-subscription.

4.3 Greenhouse Gas Mitigation Policies

The New Jersey Sustainability Greenhouse Gas (GHG) Action Plan was designed to promote the capture of landfill methane gas for power generation, thus avoiding methane emissions (methane has a

high global warming potential). A loan fund provides New Jersey companies with low-interest loans to support initiatives including development of biomass/landfill methane energy resources, and emission offsets from landfill methane and other renewable power generation could potentially be included in GHG trading systems.

5. POWER-SECTOR RESTRUCTURING POLICIES

Power sector restructuring is having a profound effect on electric power technologies, costs, prices, institutions, and regulatory frameworks. Restructuring trends are changing the traditional mission and mandates of electric utilities in complex ways, and affecting environmental, social, and political conditions. There are five key trends underway that continue to influence renewable energy development, both positively and negatively, as discussed next.

5.1 Competitive Wholesale Power Markets and Removal of Price Regulation on Generation

Power generation is usually one of the first aspects of utility systems to be deregulated. The trend is away from utilities monopolies toward open competition, where power contracts are signed between buyers and sellers in wholesale power markets. Distribution utilities and industrial customers gain more choices in obtaining wholesale power. Such markets may often begin with independent-power-producer (IPP) frameworks. As wholesale electricity becomes more of a competitive market commodity, price becomes relatively more important than other factors in determining a buyer's choice of electricity supplier.

The potential effects of competitive wholesale markets and IPPs on renewable energy are significant. Wholesale power markets allow IPPs to bypass the biases against renewables that traditional utility monopolies have had. Indeed, one of the very first major markets for renewable energy in the 1980s was in California, spurred by the PURPA legislation discussed earlier. In some countries, IPP frameworks have been explicitly enacted to support renewable energy. Examples are Sri Lanka and Thailand, where utility monopolies were broken and renewable energy IPPs can sell power to the grid. However, other effects of wholesale competition may stifle renewable energy development. As low-cost com-

bined-cycle gas turbines begin to dominate new generation, renewable energy has difficulty competing on the basis of price alone. In addition, the emergence of short-term power contracts and spot markets favor generation technologies with higher variable costs and lower capital costs, like fossil fuels, rather than capital-intensive but low-operating-cost technologies like renewables.

5.2 Self-Generation by End Users and Distributed Generation Technologies

Independent power producers may be the end users themselves rather than just dedicated generation companies. With the advent of IPP frameworks, utility buyback schemes (including net metering), and cogeneration technology options, more and more end users, from large industrial customers to small residential users, are generating their own electricity. Their self-generation offsets purchased power and they may even sell surplus power back to the grid. Traditionally, regulated monopoly utilities have enjoyed economic advantages from large power plants and increasing economies of scale. These advantages are eroding due to new distributed generation technologies that are cost competitive and even more efficient at increasingly smaller scales. In fact, newer technologies reduce investment risks and costs at smaller scales by providing modular and rapid capacity increments.

Renewable energy is well suited to self-generation but faces competition from other distributed generation technologies, especially those based on natural gas. Gas has become the fuel of choice for small self-producers because of short construction lead times, low fuel and maintenance costs, and modular small-scale technology. However, with restructuring, a host of distributed generation policies, including net metering, become possible (see the section on distributed generation policies). These policies often spur renewable energy investments. On the other hand, self-generators may be penalized by utility-wide surcharges that accompany restructuring, such as those for stranded generation assets (called nonby-passable competitive transition charges in the United States). Self-generators who use renewable energy must still pay these charges, based on the amount of electricity they *would* have purchased from the grid, even if actual grid consumption is small.

5.3 Privatization or Commercialization of Utilities

In many countries, utilities, historically government owned and operated, are becoming private for-profit entities that must act like commercial corporations. Even if utilities remain state owned, they are becoming commercialized—losing state subsidies and becoming subject to the same tax laws and accounting rules as private firms. In both cases, staffing may be reduced and management must make independent decisions on the basis of profitability.

The effects of privatization and commercialization on renewable energy are difficult to judge. The environmental effects of privatization can be positive or negative, depending on such factors as the strength of the regulatory body and the political and environmental policy situation in a country. Private utilities are more likely to focus more on costs and less on public benefits, unless specific public mandates exist. On the positive side, privatization may promote capital-intensive renewable energy by providing a new source of finance—capital from private debt and equity markets. However, the transition from public to private may shorten time horizons, increase borrowing costs, and increase requirements for high rates of return. All of these factors would limit investments in capital-intensive renewable energy projects in favor of lower-capital-cost, higher-operating-cost fossil-fuel technologies.

5.4 Unbundling of Generation, Transmission, and Distribution

Utilities have traditionally been vertically integrated, including generation, transmission, and distribution functions. Under some restructuring programs, each of these functions is being unbundled into different commercial entities, some retaining a regulated monopoly status (particularly distribution utilities) and others starting to face competition (particularly generators).

Unbundling can provide greater consumer incentives to self-generate using renewable energy. If retail tariffs are unbundled as well, so that generation, transmission, and distribution costs are separated, customers have more incentive to self-generate, thereby avoiding transmission and distribution charges. In addition, open-access transmission policies that go along with unbundling have been explicitly targeted to promote renewable energy in some countries. In India, open-access policies helped catalyze the wind industry there by allowing firms to

produce wind power in remote regions with good wind resources and then “wheel” the power over the transmission system to their own facilities or to third parties. Brazil enacted a 50% reduction of transmission wheeling fees for renewable energy producers, which has been credited with promoting a booming small hydro industry there. However, unbundling can also penalize inherently intermittent renewable energy if producers have to pay transmission charges on a per-capacity basis. That is, even when the transmission capacity is not being used (say the wind is not blowing), transmission charges must be paid, resulting in high average transmission costs per kWh.

5.5 Competitive Retail Power Markets and Green Power Sales

Competition at the retail level, the newest phenomena in power sector restructuring, means that individual consumers are free to select their power supplier from among all those operating in a given market. Competitive retail power markets have allowed the emergence of so-called green power suppliers who offer to sell renewable energy, usually at a premium. As green power sales grow, these suppliers are forced to invest in new renewable energy capacity to meet demand or buy power from other renewable energy producers. Green power markets have begun to flourish where retail competition is allowed, but often only in conjunction with other renewable energy promotion policies.

The Netherlands is perhaps the best-known example. Following restructuring in 2001, 1 million green power customers signed up within the first year. However, incentives played a role; a large tax on fossil-fuel generated electricity, from which green power sales were exempt, made green power economically competitive with conventional power. In the United States, green power markets are emerging in several states in response to state incentives and aggressive marketing campaigns by green power suppliers. At least 30 U.S. states have green pricing programs. Four states have mandatory green power policies that require utilities to offer customers opportunities to support renewable energy. California became one of the largest markets, with more than 200,000 customers, but this was aided by a 1 cent/kWh subsidy to green power, paid for by a system benefits charge (see the section on renewable energy promotion policies).

6. DISTRIBUTED GENERATION POLICIES

Distributed generation avoids some of the costs of transmission and distribution infrastructure and power losses, which together can total up to half of delivered power costs. Policies to promote distributed generation—including net metering, real-time pricing, and interconnection regulations—do not apply only to renewable energy but nevertheless can strongly influence renewable energy investments.

6.1 Net Metering

Net metering allows a two-way flow of electricity between the electricity distribution grid and customers with their own generation. When a customer consumes more power than it generates, power flows from the grid and the meter runs forward. When a customer installation generates more power than it consumes, power flows into the grid and the meter runs backward. The customer pays only for the net amount of electricity used in each billing period and is sometimes allowed to carryover net electricity generated from month to month. Net metering allows customers to receive retail prices for the excess electricity they generate at any given time. This encourages customers to invest in renewable energy because the retail price received for power is usually much greater than it would be if net metering were not allowed and customers had to sell excess power to the utility at wholesale rates or avoided costs. Electricity providers may also benefit from net metering programs, particularly with customer-sited PV, which produces electricity during peak periods. Such peak power can offset the need for new central generation and improve system load factors.

At least 38 U.S. states now have net metering laws. Size limits on net metered systems typically range from 10 kW to 100 kW, with the exception of a few states that do not limit system size or overall enrollment. Net metering is common in parts of Germany, Switzerland, and the Netherlands and allowed by at least one utility in the UK. Thailand is one of the few developing countries to have enacted net metering laws, following a pilot project for net-metered rooftop PV in the 1990s.

6.2 Real-Time Pricing

Real-time pricing, also known as dynamic pricing, is a utility rate structure in which the per-kWh charge

varies each hour based on the utility's real-time production costs. Because peaking plants are more expensive to run than baseload plants, retail electricity rates are higher during peak times than during shoulder and off-peak times under real-time pricing. When used in conjunction with net metering, customers receive higher peak rates when selling power into the grid at peak times. At off-peak times the customer is likely purchasing power from the grid, but at the lower off-peak rate. Photovoltaic power is often a good candidate for real-time pricing, especially if maximum solar radiation occurs at peak-demand times of day when power purchase prices are higher. Real-time metering equipment is necessary, which adds complexity and expense to metering hardware and administration.

Real-time pricing has been used with some large power consumers for decades. For example, power companies in Nova Scotia and New York state offer real time pricing rates for large commercial and industrial customers that vary hourly according to the varying cost of generation. In a recent pilot project, California installed 23,000 real-time meters for large customers at a cost of \$35 million. In response, summer peak demand by those customers dropped by 500 MW under time-of-use pricing, which would allow the utility to avoid \$250 million to \$300 million in capacity additions. Although real-time pricing has not become widespread, with favorable rate structures it has the potential to provide significant incentives for grid-connected renewable development.

6.3 Interconnection Regulations

Nondiscriminatory interconnection laws and regulations are needed to address a number of crucial barriers to interconnection of renewable energy with the grid. Interconnection regulations often apply to both distributed generation and remote generation with renewable energy that requires transmission access, such as wind power.

6.3.1 Legal Access

The ability to legally connect a renewable energy system to a grid depends on federal, state, and local government rules and regulations. These policies both allow connection and determine how physical connection is achieved. In the United States, the legal right to connect to the grid is provided for in federal laws such as the Public Utilities Regulatory Policies Act (PURPA) of 1978 and by state net metering statutes.

6.3.2 *Dynamic Generation and Transmission Scheduling*

Historically, transmission policies have often imposed severe penalties on unscheduled deviations from projected (advance-scheduled) power generation. These penalty structures render intermittent generation, such as wind or PV, uneconomic. Real-time accounting of power transfer deviations that provides charges or credits to producers based on the value of energy at the time of the deviation, as well as elimination of discriminatory deviation penalties, allows intermittent renewable energy to compete more equitably with traditional generation. Policies that allow near-time or real-time scheduling of the output levels of intermittent resources can further reduce deviation costs. For example, wind farms are able to predict their output much more accurately up to an hour in advance of generation and thus can be better scheduled hour by hour rather than a day ahead.

6.3.3 *Elimination of Rate Pancaking*

Because distributed renewables, such as wind, are often remotely located, they can incur high transmission fees as power crosses multiple jurisdictions to get to the customer. Such cumulative addition of transmission fees is known as rate pancaking. Elimination of access rate pancaking, either by consolidation of long-distance tariffs under a regional transmission organization or by creating access waiver agreements between multiple owner/operators can reduce discrimination against wind and other remote distributed renewables.

6.3.4 *Capacity Allocation*

When demand for a transmission path exceeds its reliable capacity, transmission congestion occurs. In such circumstances, system operators must allocate available capacity among competing users. Traditional utility policies often favor early market entrants, grandfathering them into capacity allocation rules. Wind power is particularly susceptible to transmission constraints, as it is generally located far from load. Elimination of grandfathering would allow transmission users to bid for congested capacity on an equal footing. Allowing wind to bid for congested capacity closer to the operating hour and reducing congestion through transmission line upgrades would also reduce barriers to wind energy development.

6.3.5 *Standard Interconnection Agreements*

Utilities may require the same interconnection procedures for small systems as are required for

large independent power-production facilities. The process of negotiating a power purchase/sale contract with the utility can be very expensive, and utilities can charge miscellaneous fees that greatly reduce the financial feasibility of small grid-connected renewable installations. Standardized interconnection agreements can expedite this process. Some believe that Texas provides a good model for renewable interconnection. Under a standard agreement, renewable developers pay only for the direct costs of connecting the plant to the local system, but not for upgrades to the grid necessary to carry additional capacity. This allows generators to compete more equally.

7. *RURAL ELECTRIFICATION POLICIES*

Historically, renewable energy in developing countries has come from direct donor assistance and grants for equipment purchases and demonstrations. In recent years a number of new approaches have emerged for promoting renewable energy in off-grid rural areas, including energy service concessions, private entrepreneurship, microcredit, and comparative line extension analysis.

7.1 *Rural Electrification Policy and Energy Service Concessions*

Many developing countries have explicit policies to extend electric networks to large shares of rural populations that remain unconnected to power grids (globally, an estimated 1.7 billion people). However, in many areas, full grid extension is too costly and unrealistic. Policies and rural electrification planning frameworks have recently started to emerge that designate certain geographic areas as targets for off-grid renewable energy development. These policies may also provide explicit government financial support for renewable energy in these areas. Such financial support is starting to be recognized as a competitive alternative to government subsidies for conventional grid extensions. Countries taking the lead with such policies include Argentina, China, India, Morocco, the Philippines, South Africa, and Sri Lanka.

One form this government support can take is so-called energy service concessions. With a concession, the government selects one company to exclusively serve a specific geographic region, with an obligation to serve all customers who request service. The

government also provides subsidies and regulates the fees and operations of the concession. Rural energy-service concessions may employ a mixture of energy sources to serve customers, including diesel generators, minihydro, photovoltaic, wind, and biomass. Argentina, Morocco, and South Africa have initiated policies to develop rural concessions, with ambitious targets of 200,000 rural households in South Africa and 60,000 in Argentina. But the actual experience with this approach has been limited so far to just a few thousand installations.

7.2 Rural Business Development and Microcredit

Private entrepreneurship is increasingly recognized as an important strategy to fulfill rural energy goals. Thus, rural electrification policies have begun to promote entrepreneurship. Promising approaches are emerging that support rural entrepreneurs with training, marketing, feasibility studies, business planning, management, financing, and connections to banks and community organizations. These approaches include bundling renewable energy with existing products. Bundling can reduce costs if vendors of existing products and services add renewable energy to their activities—and use their existing networks of sales outlets, dealers, and service personnel. Dealers of farm machinery, fertilizers, pumps, generators, batteries, kerosene, LPG, water, electronics, telecommunications, and other rural services can bundle renewable energy with these services.

In conjunction with entrepreneurship, consumer microcredit has emerged as an important tool for facilitating individual household purchases of renewable energy systems like solar home systems. Credit may be provided either by the system vendors themselves, by rural development banks, or by dedicated microcredit organizations. Notable examples of consumer microcredit for solar home systems have emerged in five developing countries. In Bangladesh, Grameen Shakti, a nonprofit vendor, has offered consumer credit for terms up to 3 years. The Vietnam Women's Union offered similar credit terms for systems sold by private vendors. In Sri Lanka, Sarvodaya, a national microfinance organization, has offered credit on terms up to 5 years. In Zimbabwe, vendors sold several thousand systems on credit provided by the Agricultural Finance Corporation. And in India, new forms of rural microcredit have started to emerge. By 2002, the cumulative number of

solar home system purchases made with credit in these countries had exceeded 50,000, but this was still a small fraction of the total number of solar home systems worldwide, estimated at 1.2 million.

7.3 Comparative Line Extension Analyses

Economic comparisons of line extension versus distributed renewable energy investment are also emerging in developed countries. At least four U.S. states have power line extension policies requiring that, in cases where utility customers must pay a portion of construction costs for utility power line extension to a remote location, the utility must provide information about onsite renewable energy technology options. Some of these policies require the utility to perform a cost-benefit analysis comparing line extension with off-grid renewable energy. Renewable energy options may be less expensive for rural customers, but without line extension policies, many customers would not be aware of this.

8. SUMMARY

Public support for renewable energy expanded rapidly in the late 1990s and early 2000s. A wide variety of policies are designed explicitly to promote renewable energy, while other policies focus on power sector restructuring or environmental issues more broadly and have more indirectly affected renewable energy. Experience with renewable energy policies around the world is still emerging, and more understanding is needed of the impacts of various policies. Thus, many policies could still be considered experimental in nature. Of all the policies surveyed in this article, the ones that appear to have contributed the most to renewable energy development during the 1990s and early 2000s are (1) direct equipment subsidies and rebates, net metering laws, and technical interconnection standards in the case of solar PV; (2) investment tax credits, production tax credits, and European electricity feed-in laws in the case of wind; (3) grid-access and wheeling policies supporting independent power producers and third-party sales in the case of biomass and small hydro power. Many of the trends toward restructuring of power-sector institutions and regulation that were underway throughout the 1990s have had both positive and negative influences on renewable energy. Policymakers and policy advocates have many options from which to choose and a slowly emerging body of experience and results to guide those choices.

SEE ALSO THE FOLLOWING ARTICLES

Computer Modeling of Renewable Power Systems • International Energy Law and Policy • Nongovernmental Organizations (NGOs) and Energy • Public Reaction to Renewable Energy Sources and Systems • Renewable Energy and the City • Renewable Energy in Europe • Renewable Energy in Southern Africa • Renewable Energy in the United States • Renewable Portfolio Standard

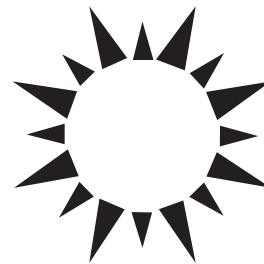
Further Reading

- Bolinger, M., Wiser, R., Milford, L., Stoddard, M., and Porter, K. (2001). States emerge as clean energy investors: A review of state support for renewable energy. *Electricity J.* (Nov.), 82–95.
- Dunn, S. (2000). “Micropower: The Next Electrical Era.” World-watch Institute, Washington, DC.
- Hirsh, R. F., and Serchuk, A. H. (1999). Power switch: Will the restructured electric utility system help the environment? *Environment* 41, 32–39. 4–9.
- International Energy Agency. (2003). “Renewable Energy Policies and Measures in IEA Countries.” Paris.
- Martinot, E., Chaurey, A., Moreira, J., Lew, D., and Wamukonya, N. (2002). Renewable energy markets in developing countries. *Ann. Rev. Energy Environment* 27, 309–348.
- Rader, N., and Hempling, S. (2001). “The Renewables Portfolio Standard: A Practical Guide.” National Association of Regulatory Utility Commissioners, Washington, DC.
- Reiche, D. (2002). “Handbook of Renewable Energy in the European Union: Case Studies of All Member States.” Verlag Peter Lang, Bern.
- Sawin, J. L. (2001). “The Role of Government in the Development and Diffusion of Renewable Energy Technologies: Wind Power in the United States, California, Denmark, and Germany, 1970–2000.” Ph. D. thesis, Tufts University, Fletcher School of Law and Diplomacy. UMI, Ann Arbor, MI.
- United Nations Environment Programme (2000). “Natural Selection: Evolving Choices for Renewable Energy Technology and Policy.” Paris.
- U.S. Energy Information Administration (2001). “Renewable Energy Annual.” U.S. Government Printing Office. Washington, DC.
- Wiser, R., Bolinger, M., Milford, L., Porter, K., and Clark, R. (2002). “Innovation, Renewable Energy, and State Investment: Case Studies of Leading Clean Energy Funds.” Lawrence Berkeley National Laboratory, Berkeley, CA.

Renewable Energy, Taxonomic Overview

DANIEL M. KAMMEN

University of California, Berkeley
Berkeley, California, United States



-
1. Introduction
 2. Biomass Energy
 3. Wind Energy
 4. Solar-Photovoltaic and Solar-Thermal Technologies
 5. Hydropower
 6. Geothermal Energy
 7. Renewable Energy System Cost and Performance
 8. Conclusions

Glossary

bioenergy The energy derived from any cellulosic plant or agricultural or animal material that is used as a fuel.

hydropower The energy derived, at a variety of scales, from water pressure; used to drive turbines.

learning curve A mathematical description of the relationship between the average price of a manufactured commodity and the market price. For a wide variety of materials, the decrease in market price is roughly 20% for each doubling of the total quantity of the item manufactured.

photovoltaics Semiconductor technology involving photon bombardment, which permits the flow of electrons.

renewable energy The energy derived from solar, wind, biomass, geothermal, or hydro sources. There is some debate about hydropower, of which the different sources are considered either to all fall into the renewable category, or are considered renewable only when involving either a small system or a “run of the river” technology.

wind energy The energy contained in the movement of air masses, generally captured by the use of mechanical blades driven by wind energy.

The potential of renewable energy sources is enormous, because, in principle, they can meet many times the world's energy demand. Renewable energy sources such as biomass, wind, solar, hydropower,

and geothermal can provide sustainable energy services, based on the use of routinely available, indigenous resources. As the costs of renewable sources decline and the price of oil and gas continue to fluctuate, a transition to renewables-based energy systems is looking increasingly likely. In the past 30 years, solar and wind power systems have experienced rapid sales growth, declining capital costs, declining costs of electricity generated, and continued improvement in performance characteristics. In fact, fossil fuel and renewable energy prices, and the social and environmental costs of each, are heading in opposite directions, and the economic and policy mechanisms needed to support the widespread dissemination and sustainable markets for renewable energy systems are rapidly evolving. It is becoming clear that future growth in the energy sector will be primarily in the new regime of renewable energy, and to some extent natural gas-based systems, not in conventional oil and coal sources. Because of these developments, market opportunity now exists to both innovate and to take advantage of emerging markets to promote renewable energy technologies, with the additional assistance of governmental and popular sentiment. The development and use of renewable energy sources can enhance diversity in energy supply markets, contribute to securing long-term sustainable energy supplies, help reduce local and global atmospheric emissions, and provide commercially attractive options to meet specific energy service needs, particularly in developing countries and rural areas, creating new employment opportunities.

1. INTRODUCTION

Conventional energy sources based on oil, coal, and natural gas have proved to be highly effective drivers

of economic progress, but at the same time damaging to the environment and to human health. Furthermore, fossil fuel-driven economies tend to be cyclical in nature, due to the effects of oligopoly in production and distribution. The traditional fossil fuel-based energy sources are facing increasing pressure on a host of environmental fronts, with perhaps the most serious challenge confronting the future use of coal being the Kyoto Protocol greenhouse gas (GHG) reduction targets. It is now clear that any effort to maintain atmospheric levels of CO₂ below even 550 ppm cannot be based fundamentally on an oil- and coal-powered global economy, barring radical carbon sequestration efforts.

The potential of renewable energy sources is enormous; they have the potential to meet many times the world's energy demand. Renewable energy sources such as biomass, wind, solar, hydropower, and geothermal can provide sustainable energy services, based on the use of routinely available, indigenous resources. A transition to renewables-intensive energy future is looking increasingly likely. The costs of solar and wind power systems have dropped substantially in the past 30 years, and continue to decline, whereas the price of oil and gas continue to fluctuate. Renewable energy prices and the social and environmental costs of using fossil fuels are heading in opposite directions. Furthermore, the economic and policy mechanisms needed to support the widespread dissemination and sustainable markets for renewable energy systems have also rapidly evolved. It is becoming clear that future growth in the energy sector is primarily in the new regime of renewable (and to some extent natural gas-based) systems, and not in conventional oil and coal sources. Financial markets are awakening to the future growth potential of renewable and other new energy technologies, and this is a likely harbinger of the economic reality of truly competitive renewable energy systems.

Renewable energy systems are usually founded on a small-scale, decentralized paradigm that is inherently conducive to, rather than at odds with, many electricity distribution, cogeneration (combined heat and power), environmental, and capital cost issues. As an alternative to custom, onsite construction of centralized power plants, renewable systems based on photovoltaic (PV) arrays, windmills, biomass, or small hydropower setups can be viewed as mass-produced "energy appliances" capable of being manufactured at low cost and tailored to meet specific energy loads and service conditions. These systems can have dramatically reduced as well as

widely dispersed environmental impacts, rather than larger, more centralized impacts that in some cases are serious contributors to ambient air pollution, acid rain, and global climate change.

Renewable energy sources currently supply somewhere between 15 and 20% of the world's total energy demand. The supply is dominated by traditional biomass, mostly fuel wood used for cooking and heating, especially in developing countries in Africa, Asia, and Latin America. A major contribution is also obtained from the use of large hydropower systems with nearly 20% of the global electricity supply being provided by this source. New renewable energy sources (solar energy, wind energy, modern bioenergy, geothermal energy, and small hydropower) are currently contributing about 2%. A number of scenario studies have investigated the potential contribution of renewables to global energy supplies, indicating that in the second half of the 21st century their contribution might range from the present figure of nearly 20% to more than 50%, with the right policies in place.

2. BIOMASS ENERGY

2.1 Introduction

Biomass is the term used for all organic material originating from plants (including algae), trees, and crops and is essentially the collection and storage of the sun's energy through photosynthesis. Biomass energy, or bioenergy, is the conversion of biomass into useful forms of energy such as heat, electricity, and liquid fuels.

Biomass for bioenergy comes either directly from the land, as dedicated energy crops, or from residues generated in the processing of crops for food or other products, such as pulp and paper from the wood industry. Another important contribution is from postconsumer residue streams, including construction and demolition wood, pallets used in transportation, and the clean fraction of municipal solid waste (MSW). The process that converts biomass to bioenergy can be considered as management of the flow of solar-generated materials, food, and fiber in our society. These interrelationships are shown in Fig. 1, which presents the various resource types and applications, showing the flow of their harvest and residues to bioenergy applications. Not all biomass is directly used to produce energy, but rather can be converted into intermediate energy carriers called biofuels. This includes charcoal (higher energy

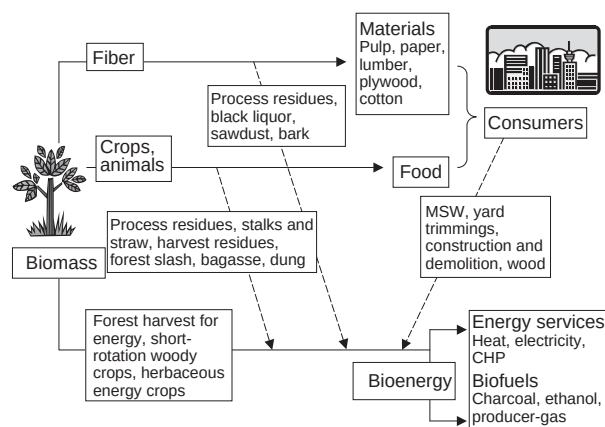


FIGURE 1 Biomass and bioenergy flowchart. MSW, Municipal solid waste; CHP, combined heat and power. From Overend (2000).

density solid fuel), ethanol (liquid fuel), or producer gas (from gasification of biomass).

Biomass was the first energy source harnessed by humans, and for nearly all of human history, wood has been our dominant energy source. Only during the past century, with the development of efficient techniques to extract and burn fossil fuels, have coal, oil, and natural gas replaced wood as the industrialized world's primary fuel. Today, some 40 to 55 exajoules ($\text{EJ} = 10^{18}$ joules) per year of biomass is used for energy, out of about 450 EJ per year of total energy use, or an estimated 10–14%, making it the fourth largest source of energy behind oil (33%), coal (21%), and natural gas (19%). The precise amount is uncertain because the majority is used noncommercially in developing countries, where biomass accounts for an estimated one-third of primary energy use. In the poorest countries, up to 90% of all energy is supplied by biomass. Over 2 billion people cook by direct combustion of biomass, and such traditional uses typically involve the inefficient use of biomass fuels, largely from low-cost sources such as natural forests, which can further contribute to human health impacts, deforestation, and environmental degradation.

The picture of biomass utilization in developing countries is sharply contrasted by that in industrialized countries. On average, biomass accounts for only 3–4% of total energy use in the latter, although where policies supportive of biomass use are in place, e.g., in Austria, Sweden, and Finland, the biomass contribution reaches 12, 18, and 23%, respectively. Most biomass in industrialized countries is converted into electricity and process heat in cogeneration systems (combined heat and power production) at

industrial sites or at municipal district heating facilities. This enables a greater variety of energy services to be derived from the biomass sources, which are much cleaner than fossil fuels, and more efficient use of the available biomass resources than is typical in developing countries.

Biomass energy has the potential to be “modernized” worldwide—that is, produced and converted efficiently and cost-competitively into more convenient forms such as gases, liquids, or electricity. A variety of technologies can convert solid biomass into clean, convenient energy carriers over a range of scales, from household/village to large industrial. Some of these technologies are commercially available today whereas others are still in the development and demonstration stages. If widely implemented, such technologies could enable biomass energy to play a much more significant role in the future than it does today, especially in developing countries.

2.2 Future Role of Biomass

Modernized biomass energy is projected to play a major role in the future global energy supply. This is being driven not so much by the depletion of fossil fuels, which has ceased to be a defining issue with the discovery of new oil and gas reserves and the large existing coal resources, but rather by the recognized threat of global climate change, caused largely by the burning of fossil fuels. The carbon neutrality of biomass energy (when produced sustainably) and its relatively even geographical distribution, coupled with the expected growth in energy demand in developing countries, where affordable alternatives are not often available, make it a promising energy source in many regions of the world for the 21st century.

Most households in developing countries that use biomass fuels today do so either because it is available at low (or zero) financial cost or because they lack access to or cannot afford higher quality fuels. As incomes rise, preferences tend to shift away from biomass. For example, in the case of cooking, consumer preferences shift with increasing income from dung to crop residues, fuel wood, coal, charcoal, kerosene, liquefied petroleum gas, natural gas, and electricity (the well-known household energy ladder). This shift away from biomass energy as incomes rise is associated with the quality of the energy carrier used, rather than with the primary energy source. If biomass energy is instead modernized, then wider use is conceivable, along with

TABLE I

Role of Biomass in Future Global Energy Use^a

Source	Time frame (year)	Total projected global energy demand (EJ/year)	Contribution of biomass to energy demand [EJ/year (percentage of total)]	Remarks
Shell Petroleum (1996)	2060	940	207 (22%)	Sustained growth ^b
	2060	1500	221 (14%)	Dematerialization ^c
Intergovernmental Panel on Climate Change (1996)	2050	560	180 (32%)	Biomass-intensive energy system development
	2100	710	325 (46%)	
Greenpeace (1993)	2050	610	114 (19%)	Fossil fuels are phased out during the 21st century
	2100	986	181 (18%)	
Johansson <i>et al.</i> (1993)	2025	395	145 (37%)	RIGES model calculation
	2050	561	206 (37%)	
World Energy Council (1993)	2050	671	94 (14%)	Outcome of one scenario
	2100	895	132 (15%)	

^aBased on five different studies. *Sources.* Hall (1998) and WEA (2000).^bBusiness-as-usual scenario.^cEnergy conservation scenario.

benefits such as reduced indoor air pollution. For example, in household cooking, compared to solid fuels, gaseous or liquid cooking fuels can be used far more efficiently and conveniently, reaching many more families and emitting far fewer toxic pollutants.

Estimates of the technical potential of biomass energy are much larger than the present world energy consumption. If agriculture is modernized up to reasonable standards in various regions of the world, several billions of hectares may be available for biomass energy production well into this century. This land would comprise degraded and unproductive lands or excess cropland, preserving the world's nature areas and quality cropland. Table I gives a summary of the potential contribution of biomass to the world's energy supply according to a number of studies and influential organizations. Although the percentile contribution of biomass varies considerably, depending on the expected future energy demand, the absolute potential contribution of biomass in the long term is high, from about 100 to 300 EJ per year.

An Intergovernmental Panel on Climate Change (IPCC) study has explored five energy supply scenarios for satisfying the world's growing demand for energy services in the 21st century while limiting cumulative CO₂ emissions between 1990 and 2100 to fewer than 500 gigatonnes (Gt) of carbon. In all scenarios, a substantial contribution from carbon-

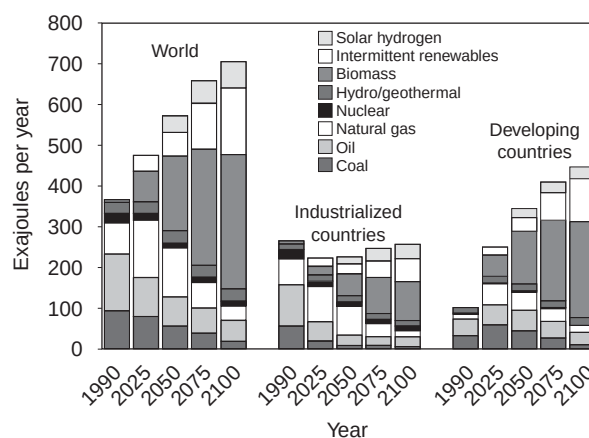


FIGURE 2 Primary commercial energy use by source for the biomass-intensive variant of the Intergovernmental Panel on Climate Change 1996 model, shown for the world, for industrialized countries, and for developing countries. From Sivan and Larson (2000).

neutral biomass energy as a fossil fuel substitute is included to help meet the CO₂ emissions targets. When biomass is grown at the same average rate as it is harvested for energy, it is approximately carbon neutral: carbon dioxide extracted from the atmosphere during growth is released back to the atmosphere during conversion to energy. Figure 2 shows the results for the IPCC's most biomass-intensive scenario, whereby biomass energy contributes

180 EJ/year to global energy supply by 2050, satisfying about one-third of the total global energy demand and about half of total energy demand in developing countries. Roughly two-thirds of the global biomass supply in 2050 is assumed to be produced on high-yield energy plantations covering nearly 400 million hectares, or an area equivalent to one-quarter of present planted agricultural area. The other one-third comes from residues produced by agricultural and industrial activities. On the other hand, the large role biomass is expected to play in future energy supplies can be explained by several considerations. First, biomass fuels can substitute more or less directly for fossil fuels in the existing energy supply infrastructure. Intermittent renewables such as wind and solar energy are more challenging to the ways we distribute and consume energy. Second, the potential resource is large. Third, in developing countries, demand for energy is rising rapidly due to population increases, urbanization, and rising living standards. Although some fuel switching occurs in the process, the total demand for biomass will also tend to increase, as is currently seen for charcoal. Consequently, there is a growing consensus that energy policies will need to be concerned about the supply and use of biofuels while supporting ways to use these fuels more efficiently and sustainably.

2.3 Biomass Energy Conversion Technologies and Applications

Generating modern energy carriers (electricity, gas, and liquid fuels) from biomass, for use at the household (~ 10 kW), community (~ 100 kW), or industrial ($\sim$ MW) scale, can be accomplished by a variety of technologies. The different technologies tend to be classed in terms of either the conversion process they use or the end product produced.

2.3.1 Combustion

The predominant technology in the world today for electricity generation from biomass, at scales above 1 MW, is the steam–Rankine cycle. This consists of direct combustion of biomass in a boiler to raise steam, which is then expanded through a turbine. The steam–Rankine technology is a mature technology introduced into commercial use about 100 years ago. The typical capacity of existing biomass power plants ranges from 1 to 50 MWe (megawatts, electricity), with an average around 20 MWe. Energy conversion efficiencies are relatively low (15–25%), due to their

small size, although technologies and processes to increase these efficiencies are being developed. Steam cycle plants are often located at industrial sites, where the waste heat from the steam turbine is recovered and used for meeting industrial process heat needs. Such combined heat and power (CHP), or cogeneration, systems provide greater levels of energy services per unit of biomass consumed than do systems that generate only power, and can reach overall efficiencies of greater than 80%.

Biomass power generating capacity grew rapidly in the United States in the 1980s largely as the result of incentives provided by the Public Utilities Regulatory Policies Act of 1978 (PURPA), which required utilities to purchase electricity from cogenerators and other qualifying independent power producers at a price equal to the utilities' avoided costs. Currently in the United States, the installed biomass-electric generating capacity is about 7 GW (not including generating capacity of ~ 2.5 GW from MSW and ~ 0.5 GW from landfill gas). The majority of this capacity is located at pulp and paper mills, where biomass fuels are available as by-products of processing. There are also a substantial number of biomass power plants in California that use agricultural processing wastes as fuel. A significant number of biomass power plants are also found in Scandinavia, especially Sweden, where carbon taxes have encouraged recent expanded use of such systems for combined district heating and power production. By comparison to the steam–Rankine power-generating capacity installed in Organization for Economic Cooperation and Development (OECD) countries, there is relatively little capacity installed in developing countries. The most significant installation of steam–Rankine capacity in developing countries is at factories making sugar and/or ethanol from sugarcane, using bagasse, the fiber residue that remains after juice extraction from sugarcane.

The costs of steam–Rankine systems vary widely, depending on the type of turbine, the type of boiler, the pressure and temperature of the steam, and other factors. An important characteristic of steam turbines and boilers is that their capital costs (per unit of capacity) are scale sensitive. This, together with the fact that biomass steam–Rankine systems are constrained to relatively small scales (due to biomass fuel transport cost limitations), typically leads to biomass steam–Rankine systems that are designed to reduce capital costs at the expense of efficiency. For example, biomass-fired systems are typically designed with much more modest steam pressure and temperature than is technologically feasible, enabling

lower grade steels to be used in boiler tubes. This lowers both the costs and the efficiency. Because there are untapped supplies of low-cost biomass feedstocks available in many regions of the world, the economics of steam-Rankine systems are probably reasonable. For example, sugarcane-processing industries and sawmills present major opportunities for steam-Rankine-based combined heat and power generation from biomass.

An alternative to the direct-fired biomass combustion technologies, and considered the nearest term low-cost option, is biomass cocombustion with fossil fuels in existing boilers. Successful demonstrations using biomass as a supplementary energy source in large high-efficiency boilers have been carried out, showing that effective biomass fuel substitution can be made in the range of 10–15% of the total energy input with minimal plant modifications and no impact on the plant efficiency and operation. This strategy is economical when the biomass fuels cost less than the fossil fuels used. For fossil fuel plant capacities greater than 100 MWe, this can mean a substantial amount of biomass and related carbon savings and emissions reductions, particularly for coal substitution.

2.3.2 Gasification

Combustible gas can be produced from biomass through a high-temperature thermochemical process. The term “gasification” commonly refers to this high-temperature thermochemical conversion and involves burning biomass without sufficient air for full combustion, but with enough air to convert the solid biomass into a gaseous fuel. The product gas, called producer-gas, consists primarily of carbon monoxide, hydrogen, carbon dioxide, and nitrogen, and has a heating value of 4 to 6 MJ per normal cubic meter (Nm^3), or 10–15% of the heating value of natural gas. The intended use of the gas and the characteristics of the particular biomass (size, texture, moisture content, etc.) determine the design and operating characteristics of the gasifier and associated equipment. After appropriate treatment, the resulting gases can be burned directly for cooking or heat supply, or can be used in secondary conversion devices such as internal combustion engines or gas turbines for producing electricity or shaft work. The systems used can scale from small to medium (5–100 kW), suitable for the cooking or lighting needs of a single family or community, up to large grid-connected power generation facilities consuming several hundred of kilograms of woody biomass per hour and producing 10–100 MW of electricity.

One technology that has generated wide interest is the biomass integrated gasification combined cycle (IGCC) technology for larger scale power and combined heat and power generation in the range of 5 to 100 MWe. An IGCC system involves sizing and drying of the feedstock, followed by thermochemical gasification to produce a combustible gas, then cooling and cleaning of the gas and combustion in a gas turbine. Steam is generated using the hot exhaust of the gas turbine to drive a steam turbine, producing additional power and/or delivering lower pressure steam for heating purposes. The cascading of a gas turbine and a steam turbine in this manner is commonly called a combined cycle. In approximate terms, the IGCC technology will enable electricity to be made at double or more the efficiency of the steam cycle, and the capital costs per installed kilowatt for commercially mature IGCC units are expected to be lower than for comparably sized steam cycles. Thus, the overall economics of biomass-based power generation are expected to be considerably better with an IGCC system than with a steam-Rankine system, especially in situations in which biomass fuel is relatively expensive.

IGCC technology is expected to be commercially available within a few years, based on current demonstration and commercialization efforts worldwide. Several of the most advanced demonstration projects are in Sweden, the United Kingdom, Italy, and Brazil. In Varnamo, Sweden, the first complete biomass-fueled IGCC system has been operating for over 1500 hours on forest residues, generating 6 MW of electricity and 9 MW of heat for the local district heating system. In Yorkshire, England, construction of the Arable Biomass Renewable Energy (ARBRE) project, an IGCC facility that will generate about 10 MW of electricity from short-rotation biomass plantations, is nearly complete. It is supported under both the THERMIE program and the United Kingdom's Non-Fossil Fuel Obligation-3. The Bioelettrica Project near Pisa, Italy, an IGCC, is also under construction and will use poplar and robinia wood chips, olive residues, grape residues, and sawdust to produce 12 MWe. At a site in Bahia, Brazil, construction began in 2000 on a Global Environment Facility (GEF)/World Bank-supported demonstration project of a 32 MW IGCC power plant using plantation-grown eucalyptus for fuel. The facility will also test the use of sugarcane bagasse. It has been estimated that if IGCC technology was applied worldwide, around 25% of all current electricity generation from sugarcane-producing countries could be produced just from their available sugarcane bagasse resource.

At the intermediate scale, producer-gas from biomass gasification can be used in modified internal combustion engines (typically diesel engines), in which it can replace 70–80% of the conventional fuel required by the engine. These smaller scale biomass gasifiers, coupled to diesel/gas internal combustion engines, operate in the 100- to 200-kWe range with efficiencies on the order of 15–25%, and have been made available commercially.

The greatest technical challenge for electricity-generating gasifier systems, at all scales, continues to be cleaning the tars and oils from the producer-gas such that the system operates efficiently, is economical, and has minimal toxic by-products and air emissions.

2.3.3 *Anaerobic Digestion*

Combustible gas can also be produced from biomass through low-temperature biological processes called anaerobic (without air) digestion. Biogas is the common name for the gas produced either in specifically designed anaerobic digesters or in landfills, by capturing the naturally produced methane. Biogas is typically about 60% methane and 40% carbon dioxide, with a heating value of about 55% that of natural gas. Almost any biomass except lignin (a major component of wood) can be converted to biogas; animal and human wastes, sewage sludge, crop residues, carbon-laden industrial processing by-products, and landfill material have all been widely used.

Anaerobic digesters generally consist of an inlet, where the organic residues and other wastes are fed into the digester tank; a tank, in which the biomass is typically heated to increase its decomposition rate and is partially converted by bacteria into biogas; and an outlet, where the biomass of the bacteria that carried out the process and the undigested material remain as sludge and can be removed. The biogas produced can be burned to provide energy for cooking and space heating or to generate electricity. Digestion has a low overall electrical efficiency (roughly 10–15%, strongly dependent on the feedstock) and is particularly suited for wet biomass materials. Direct nonenergy benefits are especially significant in this process. The effluent sludge from the digester is a concentrated nitrogen fertilizer and the pathogens in the waste are reduced or eliminated by the warm temperatures in the digester tank.

Anaerobic digestion of biomass has been demonstrated and applied commercially with success in a multitude of situations and countries. In India, biogas production from manure and wastes is applied

widely in many villages and is used for cooking and power generation. Small-scale digesters have been used most extensively in India and China. Over 1.85 million cattle-dung digesters were installed in India by the mid-1990s, but about one-third of these are not operating for a variety of reasons, primarily insufficient dung supply and difficulties with the organization of dung deliveries. A mass popularization effort in China in the 1970s led to some 7 million household-scale digesters being installed, using pig manure and human waste as feed material. Many failed to work, however, due to insufficient or improper feed characteristics or poor construction and repair techniques. Estimates were that some 3 million to 4.5 million digesters were operating in the early 1980s. Since then, research, development, and dissemination activities have focused greater attention on proper construction, operation, and maintenance of digesters. One estimate is that there were some 5 million household digesters in working condition in China as of the mid-1990s. There are in addition some 500 large-scale digesters operating at large pig farms and other agroindustrial sites, and some 24,000 digesters at urban sewage treatment plants.

Several thousand biogas digesters are also operating in other developing countries, most notably South Korea, Brazil, Thailand, and Nepal. In addition, there are an estimated 5000 digesters installed in industrialized countries, primarily at large livestock processing facilities (stockyards) and municipal sewage treatment plants. An increasing number of digesters are located at food-processing plants and other industrial facilities. Most industrial and municipal digesters are used predominantly for the environmental benefits they provide, rather than for fuel production.

2.3.4 *Liquid Biofuels*

Biofuels are produced in processes that convert biomass into more useful intermediate forms of energy. There is particular interest in converting solid biomass into liquids, which have the potential to replace petroleum-based fuels used in the transportation sector. However, adapting liquid biofuels to our present-day fuel infrastructure and engine technology has proved to be nontrivial. Only oil-producing plants, such as soybeans, palm oil trees, and oilseeds such as rapeseed, can produce compounds similar to hydrocarbon petroleum products, and these have been used to replace small amounts of biodiesel. This biodiesel has been marketed in Europe and to a lesser extent in the United States,

but it requires substantial subsidies to compete with diesel. Another family of petroleum-like liquid fuels that is produced from gasified biomass is a class of synthesized hydrocarbons called Fischer–Tropsch (F–T) liquids. The process synthesizes hydrocarbon fuels [C_{10} to C_{12} hydrocarbons (kerosene) or C_3 and C_4 hydrocarbons (liquid petroleum gas)] from carbon monoxide and hydrogen gas over iron or cobalt catalysts. F–T liquids can be used as a sulfur-free diesel or can be blended with existing diesel to reduce emissions, an environmental advantage, but have yet to be produced efficiently and economically on a large scale; research and development efforts are ongoing. In addition to use as automotive fuels, F–T liquids can potentially be used as a more efficient, cleaner cooking fuel, compared to the traditional wood fuels from which it is synthesized.

Other biofuel alternatives to petroleum-based fuels are biomass-produced alcohols, which can replace gasoline or kerosene. The most widely produced bioalcohol today is ethanol, derived from the fermentation of biomass. In industrialized countries, ethanol is most commonly produced from food crops such as corn; in the developing world, it is produced from sugarcane. Its most prevalent use is as a gasoline fuel additive to boost octane levels or to reduce dependence on imported fossil fuels. In the United States and Europe, ethanol production is still far from competitive when compared to gasoline and diesel prices, and the overall energy balance of such systems has not been very favorable. The Brazilian Proalcohol ethanol program, initiated in 1975, has been successful due to the high productivity of sugarcane, although subsidies are still required. Two other potential transportation biofuels are methanol and hydrogen. They are both produced via biomass gasification and may be used in future fuel cells.

Although ethanol production from maize and sugarcane, both agricultural crops, has become widespread and occasionally successful, it can suffer from commodity price fluctuation relative to the fuels market. Consequently, the production of ethanol from lignocellulosic biomass (such as wood, straw, and grasses) is being given serious attention. In particular, it is thought that enzymatic hydrolysis of lignocellulosic biomass will open the way to low-cost and efficient production of ethanol. The development of various hydrolysis techniques has gained attention in recent years, particularly in Sweden and the United States, but cheap and efficient hydrolysis processes are still under development and some fundamental issues need to be resolved. Once such technical barriers are surmounted and ethanol production can

be combined with efficient electricity production from unconverted wood fractions (such as the lignin), ethanol costs could come close to current gasoline prices and overall system efficiencies could go up to about 70% (low heating value). Though the technology to make this an economically viable option still does not exist, promising technologies are in the works and there are currently a number of pilot and demonstration projects.

2.4 Implementation of Biomass Energy Systems

Raw biomass has several disadvantages as an energy source. It is bulky, has a low energy density, and direct combustion is generally highly inefficient (other than advanced domestic heaters), producing high levels of indoor and outdoor air pollution. The goal of modernized biomass energy is to increase the fuel's energy density while decreasing its emissions during production and use. Modernizing biomass energy production, however, faces a variety of challenges that must be adequately addressed and dealt with before the widespread implementation of bioenergy systems can occur. These issues include technical problems (just discussed), resource availability, environmental impacts, and economic feasibility.

2.4.1 Biomass Resources

Biomass resources are potentially the largest renewable global energy source; there is an annual primary production of around 4500 EJ with a bioenergy potential on the order of 2900 EJ, of which about 270 EJ could be considered available on a sustainable basis. The challenge is not the availability so much as the sustainable management and conversion and delivery to the consumer in the form of modern and affordable energy services. Most of the biomass used today is either a residue in a bioprocessing industry or is an opportunity fuel that is used in households for daily living needs. It is argued that if biomass is to become a major fuel in the world, as is being proposed in future energy scenarios, then residues will not suffice and energy plantations may need to supply up to 80% of the future feedstock.

The solar energy conversion efficiency of plants is low, often less than 1%. Consequently, relatively large land surfaces are required to produce a substantial amount of energy. Moreover, biomass has a low energy density (only 8 GJ/tonne of wood, 50% moisture content). For comparison, other fuels have the following energy densities: coal, 28 GJ/tonne; mineral

oil, 42 GJ/tonne; and liquefied natural gas, 52 GJ/tonne. Consequently, transportation becomes an essential element of biomass energy systems, with transportation distances becoming a limiting factor, both from an economic and energetic point of view. Although generally it has been found that for woody biomass the energy output is 10–30 times greater than the energy input necessary for fuel production and transport, the issue is less clear for the production of liquid fuels, except for ethanol from sugarcane, which does have high net energy yields.

At present, the production of biomass residues and wastes globally, including by-products of food, fiber, and forest production, exceeds 110 EJ/year, perhaps 10% of which is used for energy. Residues concentrated at industrial sites are currently the largest commercially used biomass source. Residues are not, however, always accessible for energy use. In some cases, collection and transport costs are prohibitive; in other cases, agronomic considerations dictate that residues be recycled to the land. In still other cases, there are competing nonenergy uses for residues (e.g., fodder, construction material, or industrial feedstock).

Residues are an especially important potential biomass energy source in densely populated regions, where much of the land is used for food production. In fact, biomass residues play important roles in such regions precisely because the regions produce so much food: crop production can generate large quantities of by-product residues. For example, in 1996, China generated crop residues in the field (mostly corn stover, rice straw, and wheat straw) plus agricultural processing residues (mostly rice husks, corncobs, and bagasse) totaling about 790 million tonnes, with a corresponding energy content of about 11 EJ. To put this in perspective, if half of this resource were to be used for generating electricity at an efficiency of 25% (achievable at small scales today), the resulting electricity generation would be about half of the total electricity generated from coal in China in 1996. There is also a significant potential for providing biomass for energy by growing crops specifically for that purpose. The IPCC's biomass-intensive future energy supply scenario discussed previously includes 385 million hectares of biomass energy plantations globally in 2050 (equivalent to about one-quarter of present planted agricultural area), with three-quarters of this area established in developing countries. Such levels of land use for bioenergy raise the issue of intensified competition with other important land uses, especially food production.

There are two approaches to producing energy crops. These include devoting an area exclusively to production of such crops, and commingling the production of energy and nonenergy crops, either on the same piece of land (agroforestry) or on adjacent pieces of land (farm forestry). Because energy crops typically require several years to grow before the first harvest, the second approach has the benefit of providing the energy-crop farmer with revenue from the land between harvests of energy crops. In Sweden, productive heat power generation from willow plantations has been successful, and there has also been experience in small-scale fuel wood production in India, China, and elsewhere. In Brazil, farm forestry activities have involved small farmers in the high-yield production of biomass feedstocks.

2.4.2 Environmental Impacts and Benefits

In general, renewable forms of energy are considered “green” because they cause little depletion of Earth's resources, have beneficial environmental impacts, and cause negligible emissions during power generation. Yet, although biomass is in principle renewable and can have positive environmental impacts if managed properly, it also shares many characteristics with fossil fuels, both good and bad. Biomass can be transported and stored, allowing for heat and power generation on demand, but modernized bioenergy systems can also have negative environmental impacts associated both with the growing of the biomass and with its conversion to energy.

Many bioenergy conversion technologies offer flexibility in choice of feedstock and the manner in which it is produced. In contrast, most agricultural products are subject to rigorous consumer demands in terms of taste, nutritional content, uniformity, etc. The flexibility afforded by biomass technologies makes it easier to meet the simultaneous challenges of producing biomass energy feedstocks and meeting environmental objectives. For example, unlike the case with food crops, there are good possibilities for bioenergy crops to be used to revegetate barren land, to reclaim water-logged or salinated soils, and to stabilize erosion-prone land. Biomass energy feedstocks, when properly managed, can both provide habitat and improve biodiversity on previously degraded land.

Erosion and removal of soil nutrients are problems related to the cultivation of annual crops in many regions of the world. Although, relative to a healthy natural ecosystem, bioenergy systems may increase erosion and deplete soil nutrients and quality, bioenergy production on degraded or

erosion-prone lands can instead help stabilize soils, improve their fertility, and reduce erosion. Perennial energy crops (unlike food crops) improve land cover and form an extensive root system, adding to the organic matter content of the soil. Also, removal of soil during energy crop harvest can be kept to a minimum because roots are left in place, and twigs and leaves can be left to decompose in the field, enhancing the soil's nutrient store. This helps prevent diseases and improves soil fertility and quality. Environmental benefits of biomass crops, for carbon sequestration, biodiversity, landscape enhancement, and soil stabilization, can be particularly significant if plantations are established on intensively managed agricultural land. Although energy crops can be harvested by coppicing every 3 or 4 years, the stools (rootstocks) can survive for many decades, or even centuries, becoming significant carbon sinks. In addition, there are considerable benefits to both the landscape and to biodiversity when native species are used. For example, in Europe it would be preferable to grow willows and poplars rather than eucalyptus. Willows in particular support a high biomass and diversity of phytophagous insects, which in turn can support an important food web with many bird species. Also, when feasible, the recycling of ashes from the biomass combustion can return crucial trace elements and phosphates to the soil. This is already common practice in countries such as Sweden and Austria, where some of the ashes are returned to the forest floors, and in Brazil, where stillage, a nutrient-rich remainder of sugar cane fermentation, is returned to sugar cane plantations.

Another important potential impact from bioenergy feedstock production is the introduction of agricultural inputs such as fertilizers and pesticides into the environment. Fertilizers and pesticides can adversely affect the human health, water quality, and plant and animal life. Specific effects strongly depend on the type of chemical, the quantities used, and the method of application. Current experience with perennial crops (e.g., willow, poplar, or eucalyptus) for feedstock production suggests that necessary agricultural practices meet very strict environmental standards. Compared to food crops such as cereals, application rates of agrochemicals per hectare are a factor of 5 to 20 times lower for perennial energy crops. The abundant use of fertilizers and manure in agriculture has led to considerable environmental problems in various regions in the world: nitrification of groundwater, saturation of soils with phosphates (leading to eutrophication), and problems in meeting drinking water standards. Also, the applica-

tion of phosphates has led to increased heavy metal flux to the soil. Compared to conventional agriculture, energy farming with short-rotation forestry and perennial grasses requires less fertilizer. With perennials, better recycling of nutrients is obtained. The leaching of nitrogen relating to willow cultivation can be about a factor of 2 to 10 times less than for food crops and is able to meet stringent standards for groundwater protection.

Possibly the biggest concern, and often considered the most limiting factor to the spread of bioenergy crops, is the demand on available water supplies, particularly in arid or semiarid regions. The choice of a certain energy crop can have a considerable effect on its water-use efficiency. Certain eucalyptus species, for example, have very good water-use efficiency when the amount of water needed per tonne of biomass produced is considered. However, a eucalyptus plantation on a large area could increase the local demand on groundwater and thus affect the groundwater level. On the other hand, energy crops on previously degraded land will improve land cover, which generally has positive effects on water retention and microclimate conditions. Impacts on the hydrological situation, therefore, always need to be evaluated on local level.

Finally, there is the issue of biodiversity and landscape. Biomass plantations are frequently criticized because the range of biological species they support is much narrower than is found in natural ecosystems. Although generally true, this is not always relevant. It would be relevant if a virgin forest were to be replaced by a biomass plantation—a situation that would be undesirable. However, when plantations are established on degraded lands or on excess agricultural lands, as is intended to be the case, the restored lands are very likely to support a more diverse ecology compared to the original situation. The restoration of such land is generally desirable for purposes of water retention, erosion prevention, and (micro) climate control. Furthermore, a good plantation design, including areas set aside for native flora and fauna, fitting into the landscape in a natural way, can avoid the problems normally associated with monocultures. The presence of natural predators (e.g., insects) can prevent the outbreak of pests and diseases. This issue needs more research, taking into account specific local conditions, species, and cultural aspects.

2.4.3 Economic and Production Issues

A number of key areas can be identified as being essential for the successful development and

implementation of sustainable, economically competitive bioenergy systems. The main consideration—a possible barrier—is whether the energy carriers produced are competitive. This is particularly true when specially produced biomass is used. In many situations, cheap or negative-cost biomass wastes and residues are available, thus the utilization of biomass is or could be competitive and future technology development should help further reduce the costs of bioenergy. In Sweden and Denmark, where a carbon and energy tax has been introduced, more expensive wood fuels and straw are now being used on a large scale. However, on a worldwide basis, the commercial production of energy crops is almost nonexistent. A major exception is Brazil, where subsidies have been introduced to make ethanol from sugarcane competitive with gasoline.

Closely related to the cost issue are the availability and the full-scale demonstration of advanced conversion technology that combines a high-efficiency and an environmentally sound performance with low investment costs. This is essential for competition with fossil fuels when relatively high-cost energy crops are used as energy sources. Advances in the combustion and cocombustion of biomass can considerably increase the attractiveness of combustion as a conversion technology. However, the development and the application of the IGCC technology have the potential to attain higher conversion efficiency at lower costs. Demonstration and commercialization of this technology are therefore important.

Experience with dedicated fuel supply systems based on “new” energy crops such as perennial grasses and short-rotation crops (SRCs) are very limited compared to the experience of traditional food crop cultivation and forestry techniques. Improvement of yields, increased pest resistance, management techniques, reduction of inputs, and further development of machinery are all necessary to lower costs and raise productivity. The same is true for harvesting, storage, and supply logistics. Bioenergy systems are complex in terms of organization and the number of actors that can be involved in a total energy system. The biomass is most likely to be produced by farmers or foresters whereas transport and storage are likely to be the responsibility of another party, and utilities may be responsible for the energy production. The combination of the utilities on the one hand and the agricultural system on the other will create a number of nontechnical barriers that have to be dealt with for any future system to work.

2.5 Conclusions

Biomass is one of the renewable energy sources capable of making a large contribution to the world's future energy supply. Land availability for biomass production should not be a bottleneck, provided it is combined with modernization of conventional agricultural production. Recent evaluations indicate that even if land surfaces of 400 million–700 million hectares were used for biomass production for energy about halfway the next century, this could be done without conflicting with other land-use functions and nature preservation. This can be obtained in part by better agricultural practices and in part by making use of huge areas of unproductive degraded lands. Latin America, Africa, Asia, and to a lesser extent Eastern Europe and North America represent a large potential for biomass production.

The types of biomass that can be used for energy are diverse, and optimal resources, technologies, and entire systems will be shaped by local conditions, both physical and socioeconomic. Perennial crops in particular may offer cheap and productive biomass production systems with low or positive environmental impacts. Technical improvements and optimized production systems along with multifunctional land use could bring biomass close to the costs of fossil fuels.

A key issue for bioenergy is that its use must be modernized to fit into a sustainable development. Conversion of biomass to energy carriers such as electricity and transportation fuels will give biomass a commercial value and will provide income for local rural economies. In order to obtain such a situation, it is essential that biomass markets and necessary infrastructure are developed, that key conversion technologies such as IGCC technology and advanced fuel production systems for methanol, hydrogen, and ethanol are demonstrated and commercialized, and that much more experience is gained with biomass production systems in a wide variety of contexts. Although the actual role of bioenergy will depend on its competitiveness versus fossil fuels and agricultural policies, it seems realistic to expect that the current contribution of bioenergy will increase during this century.

3. WIND ENERGY

3.1 Introduction

Wind has considerable potential as a global clean energy source, being both widely available, though

diffuse, and producing no pollution during power generation. Wind energy has been one of humanity's primary energy sources for transporting goods, milling grain, and pumping water for several millennia. From windmills used in China, India, and Persia over 2000 years ago to the generation of electricity in the early 20th century in Europe and North America, wind energy has played an important part in our recorded history. As industrialization took place in Europe and then in North America, wind power generation declined, first gradually as the use of petroleum and coal, both cheaper and more reliable energy sources, became widespread, and then more sharply as power transmission lines were extended into most rural areas of industrialized countries. The oil crises of the 1970s, however, triggered renewed interest in wind energy technology for grid-connected electricity production, water pumping, and power supply in remote areas, promoting the industry's rebirth.

This impetus prompted countries; notably Denmark and the United States, to establish government research and development programs to improve wind turbine technology. In conjunction with private industry research, this led to a reemergence in the 1980s of wind energy in the United States and Europe, when the first modern grid-connected wind turbines were installed. In the 1990s, this development accelerated, with wind becoming the fastest growing energy technology in the world, developing into a commercially competitive global power generation industry. In 1990, only about 2000 MW of grid-connected wind power was in operation worldwide, whereas by 1999, this figure had surpassed 10,000 MW, not including the over 1 million water-pumping wind turbines located in remote areas. Since 1990, the average annual growth rate in world wind-generating capacity has been over 24%/year, with rates of over 30% in the past 2 years. Today, there is more than 31,000 MW of installed wind power, double the capacity that was in place just 3 years earlier (Fig. 3). More than 6000 MW of new wind energy capacity was installed in 2002. This dramatic growth rate in wind power has created one of the most rapidly expanding industries in the world, with sales of roughly \$7 billion in 2002, and predictions of 10-fold growth over the next decade. Most 2000 forecasts for installed capacity were quickly eclipsed, and wind power passed the 10,000-MW mark in early 1999 and is now three times that level.

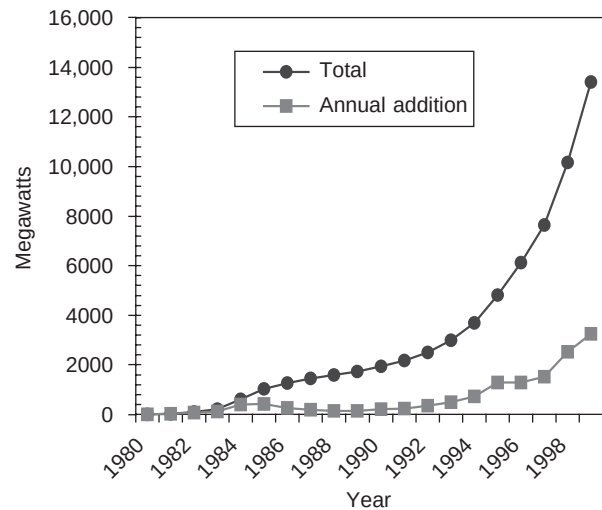


FIGURE 3 World wind-generating capacity, total and annual addition. From American Wind Energy Association (2003).

3.2 Economics of Wind Energy

Larger turbines, more efficient manufacturing, and careful siting of wind machines have brought wind power costs down precipitously, from \$2600/kW in 1981 to \$800/kW in 1998 and still lower today. New wind farms in some areas have now reached economic parity, or have beaten, new coal-based power plants, and as the technology continues to improve, further cost declines are projected, which could make wind power the most economical source of electricity in some countries. Market growth, particularly in Europe, has been stimulated by a combination of favorable governmental policies, lower costs, improved technology, and concern over environmental impacts of energy use. Compared to wind turbines built in 1981, modern turbines generate 56 times the energy at only nine times the cost.

Wind energy is currently one of the most cost-competitive renewable energy technologies. Worldwide, the cost of generating electricity from wind has fallen by more than 80%, from about 38¢/kWh (U.S.) in the early 1980s to a current range of 3–6¢/kWh (U.S.) levelized over a plant's lifetime, and analysts forecast that costs will drop an additional 20–30% in the next 5 years. Consequently, in the not-too-distant future, analysts believe that wind energy costs could fall lower than most conventional fossil fuel generators, reaching a cost of 2.5¢/kWh (U.S.) (Fig. 4). Wind technology does not have fuel requirements, as do coal, gas, and petroleum technologies. However, both the equipment costs and the costs of accommodating special characteristics such as intermittence, resource variability,

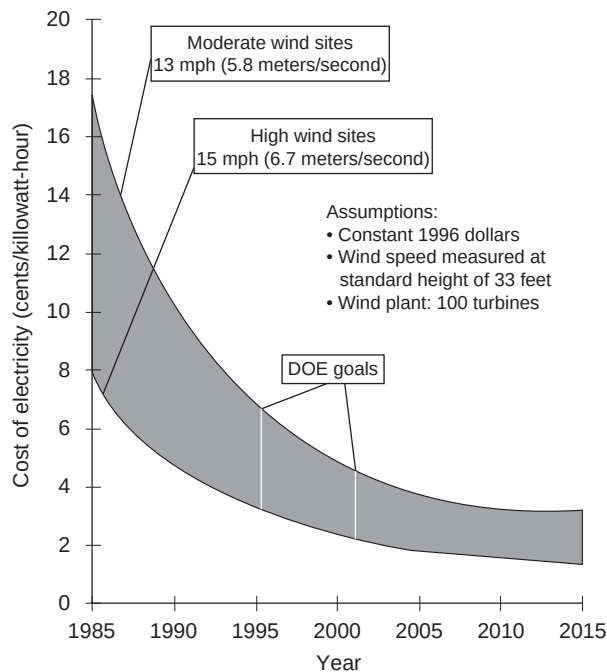


FIGURE 4 Trends in wind energy costs. From the U.S. Department of Energy (1998), (see the DOE Web site at <http://www.eren.doe.gov>).

competing demands for land use, and transmission and distribution availability can add substantially to the costs of generating electricity from wind.

3.3 Potential for Wind Energy: Technical, Resource, and Environmental Issues

The main technical parameter determining the economic success of a wind turbine system is its annual energy output, which in turn is determined by parameters such as average wind speed, statistical wind speed distribution, distribution of occurring wind directions, turbulence intensities, and roughness of the surrounding terrain. Of these, the most important and sensitive parameter is the wind speed (where the power in the wind is proportional to the third power of the momentary wind speed), which increases with height above the ground. As a result, vertical-axis wind turbines have mostly been abandoned in favor of the taller traditional horizontal-axis configuration. As accurate meteorological measurements and wind energy maps become more commonly available, wind project developers are able to more reliably assess the long-term economic performance of wind farms.

Some of the problems with wind power involve siting wind turbines. In densely populated countries

where the best sites on land are occupied, there is increasing public resistance, making it impossible to realize projects at acceptable cost. This is one of the main reasons that countries such as Denmark and The Netherlands are concentrating on offshore projects, despite the fact that technically and economically they are expected to be less favorable, compared to good land sites. On the other hand, in countries such as the United Kingdom and Sweden, offshore projects are being planned not due to scarcity of suitable land sites, but because preserving the landscape is such an important national value. Another obstacle can be that the best wind site locations are not in close proximity to populations with the greatest energy needs, as in the U.S. midwest, making such sites impractical due to the high cost of transmission of power over long distances.

There has been a gradual growth of the unit size of commercial machines since the mid-1970s, when the typical size of a wind turbine was 30 kW installed power. By 1998, the largest units installed had a capacity of 1.65 MW, and turbines with an installed power of 2 MW are now being introduced on the market, with 3.6-MW machines now slated for construction. The trend toward larger machines is driven by the demand side of the market to utilize economy of scale, to reduce the visual impact on the landscape per unit of installed power, and to support the expectation that the offshore potential will be developed soon. Recent technical advances have also made wind turbines more controllable and grid compatible and have reduced their number of components, making them more reliable and robust.

Wind energy, although considered an environmentally sound energy option, does have several negative environmental aspects connected to its use. These include acoustic noise emission, visual impact on the landscape, impact on bird life, shadow caused by the rotor, and electromagnetic interference (influencing the reception of radio, television, and radar signals). In practice, the noise and visual impacts appear to cause the most problems for siting projects. Noise issues have been reduced by progress in aeroacoustic research, providing design tools and blade configurations that have successfully made blades considerably quieter. The impact on bird life appears to be a relatively minor problem. For instance, a research project in The Netherlands showed that bird casualties as a result of collisions with rotating rotor blades for a wind farm of 1000 MW represent only a very small fraction of birds killed from hunting, high-voltage lines, and traffic; the estimates predict a maximum level of six to seven bird

TABLE II

Estimates of Annual Total Human-Related Bird Mortality in The Netherlands^a

Causes of death	Bird victims (× 100,000)
Road kills	20–80
Power lines	10–20
Hunting	6.5
Per 1000 MW of wind power	2.1–4.6

^a Source. National Wind Coordinating Committee, Proceedings of National Avian–Wind Power Planning Meeting (1994).

collisions/turbine/year (see Table II). Avoiding endangered species habitats and major migration routes in the siting of wind farms can, for the most, part eliminate this problem.

In addition to being cost-competitive and environmentally sound, wind energy has several additional advantages over conventional fossil fuel power plants and even other renewable energy sources. First, it is modular: that is, the generating capacity of wind farms can easily be expanded because new turbines can be quickly manufactured and installed, not true for either coal-fired or nuclear power plants. Furthermore, a repair to one wind turbine does not affect the power production of all the others. Second, the energy generated by wind turbines can pay for the materials used to make them in as short as 3–4 months for good wind sites. Third, during normal operation, they produce no emissions. One estimate of wind energy potential to reduce CO₂ emissions predicts that a 10% contribution of wind energy to the world's electricity demand by 2025 would prevent the emission of 1.4 Gton/year of CO₂. Finally, there is also a strong and growing market for small wind turbines (under 100 kW), of which the United States is a leading manufacturer. Four very active U.S. manufacturers are estimated to cover a 30% market share worldwide. Small-scale turbines especially can play a significant role in rural and remote locations, particularly in developing countries, where access to the grid is either unlikely or extremely expensive.

3.4 Selected Country Profiles and Government Incentives to Promote Wind Energy

Incentives have long been viewed as a means of supporting technological developments until a new

technology becomes cost-competitive. Wind-based electricity is not yet generally competitive with alternate sources of electricity such as fossil fuels. Thus, it is still dependent on nonmarket support for development to take place. Where sufficient support has been made available, wind capacity has expanded. When support has been withdrawn, or is uncertain, wind energy development has substantially slowed. The countries that have been fueling wind energy's growth throughout this decade have mainly been in the Northern Hemisphere, in particular Europe, where issues regarding the environment, fuel security, and electricity-generating diversity are a priority. Of the 10 countries with the highest installed capacity at the end of 2000, only the United States, India, and China lie outside Europe (Table III).

The stimulus for European wind energy growth continues to be regional and national policy incentives, which have facilitated the birth and development of the wind energy industry. Collectively, high energy prices, renewable energy subsidies that include investment and other fiscal incentives, and mandated purchases that include price support systems are some of the most important parameters that have encouraged wind energy development in the European Union (EU). The heightened sensitivity that Europeans have for the environment has been a determining factor in setting such policies. Another important factor, although it is less frequently mentioned, is fuel security. Developing an indigenous resource helps diminish the dependence on imported energy sources, which is of strategic importance to many countries in the EU. According to a 2003 American Wind Energy Association (AWEA) report, Europe continued to drive the engine of global wind power growth. A total of 5871 MW (worth \$6.3 billion) was installed in the EU countries in 2002. Regional wind power capacity grew 33% to 23,056 MW. This volume of production offsets 20 million tons of coal burned annually in a conventional power plant, according to the European Wind Energy Association. Germany, Spain, and Denmark alone accounted for 89% of the wind power capacity installed in Europe in 2002. Non-EU European countries (such as Norway and Poland) contributed an additional 112 MW to the region's wind power growth.

3.4.1 United States

In the early 1980s, the United States accounted for 95% of the world's installed wind energy capacity;

TABLE III

Installed Wind Power in 1997 and 1998^a

Region	Installed MW (1997)	Cumulative MW (1997)	Installed MW (1998)	Cumulative MW (1998)
United States	29	1611	577	2141
Canada	4	26	57	83
Mexico	0	2	0	2
South and Central America	10	42	24	66
Total America	43	1681	658	2292
Denmark	285	1116	310	1420
Finland	5	12	6	18
France	8	13	8	21
Germany	533	2081	793	2874
Greece	0	29	26	55
Ireland	42	53	11	64
Italy	33	103	94	197
Netherlands	44	329	50	379
Portugal	20	39	13	51
Spain	262	512	368	880
Sweden	19	122	54	176
United Kingdom	55	328	10	338
Other European countries, including Belgium, Czech Republic Slovakia, Norway, Austria, Turkey, Switzerland, and East European countries	13	57	23	80
Total Europe	1318	4793	1,766	6553
People's Republic of China	67	146	54	200
India	120	940	82	992
Remaining Asian countries, including Korea, Japan, Malaysia, Indonesia, Thailand, and Vietnam	9	22	11	33
Total Asia	196	1108	147	1224
Australia and New Zealand	2	8	26	34
Pacific Islands	0	3	0	3
North Africa, including Egypt, Ethiopia, Libya, Tunis, Algeria, Cape Verde, and Morocco	0	9	0	9
Middle East, including Jordan, Syria, Israel, Saudi Arabia, Iran, and Iraq (excluding Egypt)	8	18	0	18
Former Soviet Union countries	1	19	1	19
Total other continents and areas	11	57	27	83
Annual MW installed capacity	1568		2597	
Cumulative World MW installed		7639		10,153

^aSource. WEA (2000). Note. The cumulative installed capacity by the end of 1998 is not always equal to the 1997 data plus installed capacity during 1998. Adjustments have been made by considering the amount of decommissioned/dismantled capacity.

the U.S. share dropped to 22% in 2002 and has essentially stagnated while other countries have increased their capacity. In 2001, 1600 MW of new capacity was installed in the United States. In 2002, that figure dropped to just over 4000 MW, although there are signs of steady growth resulting from the federal production tax credit [currently 1.7¢/kWh (U.S.) for utility-scale projects].

3.4.2 Germany

In contrast to the United States, Germany has increased its wind energy production consistently and dramatically throughout the past decade. In 2002, Germany installed over 3200 MW, or 55% of the total global new capacity. Germany overtook the United States in 1997 as the country with the highest installed wind capacity in the world, and Spain

followed suit in 2001 with a surge of new installations. A mixture of government incentives, including the Electric Feed-in Law, guaranteeing that renewable energy producers are paid up to 90% of the current domestic retail electricity price, and low interest loans in several regions, have helped make Germany the strongest wind market in the world. Germany passed another law that aims to double renewable energy's share of electricity to 10% in the next decade. This took effect in April 2000, replacing the Electric Feed-in law, and it provides a fixed level of support for all green power.

3.4.3 Denmark

Denmark ranks as the world's largest manufacturer and exporter of wind turbines, manufacturing over 60% of the world's turbines. They are 5 years ahead of their goal to produce 10% of their electricity from wind, and their new goal is to supply 50% of electric generation from wind by 2030, in part through a 4000-MW offshore wind initiative. The Danish government has been a consistent, strong supporter of its wind industry and, similar to the German legislation, their Windmill Law requires electric utilities to purchase output from private wind turbines at 85% of the consumer price of electricity, plus giving an ecotax break of 9¢/kWh (U.S.). Due to policies encouraging ownership of equipment, privately owned wind turbines represent 80% of the market. In addition, the electric utilities also receive a 1.5¢/kWh (U.S.) production subsidy for wind power generation. As of 2000, the wind energy market has been transformed from a fixed-price system to a renewable portfolio standard (RPS) such that by 2003, 20% of the country's electricity consumption will have to come from renewable sources; the current amount is between 12 and 14%. The plan is to achieve this through the trading of "green certificates" issued to renewable energy generators in proportion to their energy production. The United States is adopting RPS legislation on a state-by-state basis; the most aggressive standard (to date) is in Nevada, with a 20% RPS plan for 2015.

3.4.4 Developing Countries

Finally, among the developing countries, India continues to be the leader, followed by China. India had one of the fastest growing wind markets in the mid-1990s; growth has since slowed substantially due to underperformance of projects (often due to poor siting), transmission problems, and political and economic instability, all of which have affected investments in new projects. Extensive use of

investment tax credits has also led to unsustainable revenue losses for the government. During the 1990s, electricity consumption more than doubled in India, and despite huge expenditures in the power industry, the country still has a 12% electricity deficit and 20% peak power shortage. Due to these conditions and the continued fast rate of growth in consumption, India continues to be considered a key market for wind developers. China remains mostly an untapped market with a huge future potential for wind energy; to date, relatively little money has been invested into projects, which remain small and donor based. It is expected that major commercial development, possibly to take place in the next decade, will be needed for strong wind energy development to occur in China.

3.5 Conclusions

For the first time, we are seeing one of the emerging renewable energy generating options—wind power—in a position to compete with the generation technologies of the past century. A variety of players are engaged in pushing forward wind projects worldwide. Enron Wind Corporation acquired German turbine manufacturer Tacke, NEG Micon of Denmark has built manufacturing facilities in the United States, Vestas of Denmark has built factories in Spain and India, and many manufacturers have developed joint ventures in various countries around the world. This globalization trend is likely to continue because financial institutions are beginning to view the wind industry as a promising investment opportunity. As more countries are added to the wind energy roster, uneven development focusing on a half-dozen key markets will most likely be replaced by more balanced growth. During the next couple of years, large-scale projects are expected to be developed in Egypt, Nicaragua, Costa Rica, Brazil, Turkey, Philippines, and several other countries, totaling thousands of megawatts of newly installed capacity, and expanding the number of countries using their wind resources.

4. SOLAR-PHOTOVOLTAIC AND SOLAR-THERMAL TECHNOLOGIES

There are two basic categories of technologies that convert sunlight into useful forms of energy, aside from biomass-based systems, which do this in a broader sense by using photosynthesis from plants as

an intermediate step. First, solar-photovoltaic (PV) modules convert sunlight directly into electricity. Second, solar-thermal power systems use focused solar radiation to produce steam, which is then used to turn a turbine, producing electricity. The following discussion provides a brief overview of these technologies, along with their current commercial status.

4.1 Solar Photovoltaics

At present, annual global PV module production is over 400 MW, which translates into a more than \$2 billion/year business. In addition to the ongoing use of PV technologies in space, their present-day cost and performance also make them suitable for many grid-isolated and even grid-connected applications in both developed and developing parts of the world. PV technologies are potentially so useful that, as their comparatively high initial cost is brought down another order of magnitude, it is very easy to imagine them becoming nearly ubiquitous late in the 21st century. PV systems would then likely be employed on many scales in vastly differing environments, from microscopic cells to 100-MW or larger “central station” generating plants covering many square kilometers on Earth and in space. The technical and economic driving forces that favor the use of PV technologies in these widely diverse applications will be equally diverse. However, common among them will be the durability, high efficiency, quiet operation, and lack of moving parts that PV systems offer, and the fact that these attributes combine to provide a power source with minimum maintenance and unmatched reliability. PV system cost and performance have been steadily improving in recent years. PV manufacturing costs have fallen from about \$30/W in 1976 to \$5.00–8.00/W today, depending on the level of solar insolation at the site and other factors. These installed system costs are expected by some analysts to reach a range of \$2.00–5.00/W by 2010, and if this is achieved, PV systems could achieve a sales level of over 1600 MW/year by that time (see Fig. 5).

PV efficiency has also increased markedly over the past few decades, as shown in Fig. 6. The best thin film cells tested in laboratories in 1980 achieved an efficiency level of about 10%. This was improved to about 13% by 1990 and to over 17% in recent years, for the best thin film cells made from copper indium diselenide (CuInSe_2) or cadmium telluride (CdTe). However, the modules that are currently being manufactured have somewhat lower efficiencies compared to the recent test cells, with typical values

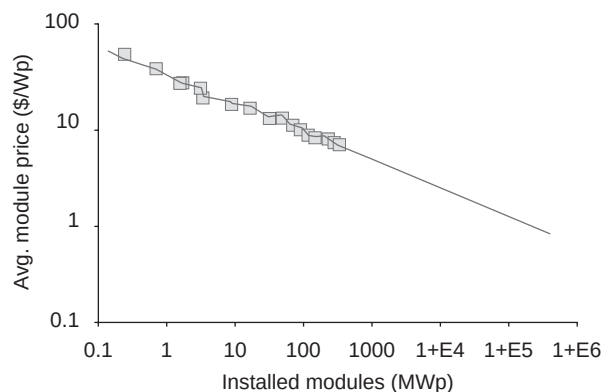


FIGURE 5 Learning curve for photovoltaics, 1976–2000. The average module price is in 1992 dollars (per watt, photovoltaic; Wp).

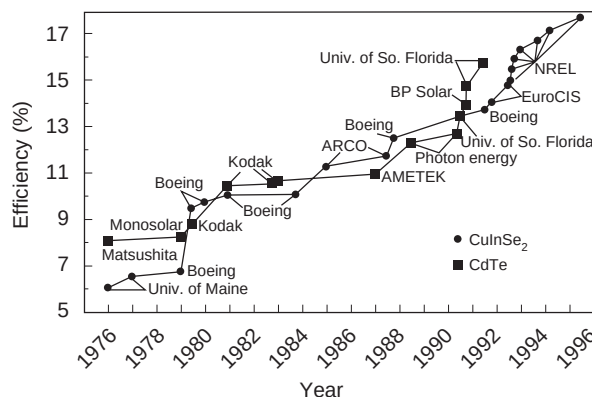


FIGURE 6 Progress in polycrystalline thin-film laboratory cell efficiencies. From the U.S. Department of Energy (1997).

in the range of 9–11% for CuInSe_2 and CdTe modules, and 8–10% for the more common amorphous silicon modules. The efficiency gains made over the past decade at the cell level will slowly translate into higher efficiency manufactured products, and this should combine with cell and module manufacturing cost reductions to aid in producing the reductions in per-watt installed system costs.

Figure 7 shows PV shipments by region and for the entire world, from 1976 to 1998. As can be seen, PV sales are growing rapidly; in fact, the sales growth rate throughout the 1990s was approximately 30%/year. Thus, PV systems are declining in cost, improving in efficiency, and increasing rapidly in sales. The cost of electricity produced from PV systems is still higher than it is from most other competing technologies, at about 30¢/kWh, but these costs are expected to continue to decline steadily. In fact, the U.S. Department of Energy projects that the cost of generating electricity from

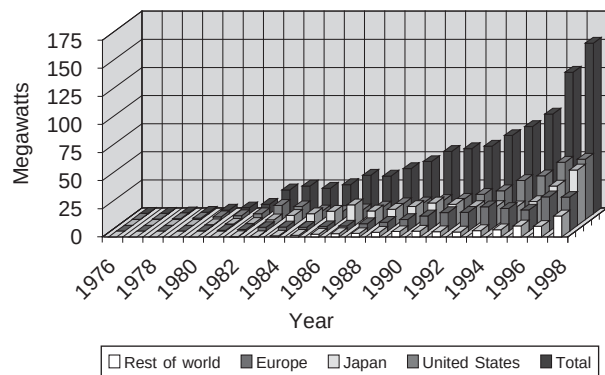


FIGURE 7 Regional and global photovoltaic shipments to different regions worldwide, 1976–2000. In 2002 (not shown), shipments reached 400 MW. From Maycock (1998).

utility-scale PV systems could be below 10¢/kWh (U.S.) by 2010, perhaps falling to about 6¢/kWh by 2020. Meanwhile, costs of electricity from residential PV systems could reach about 18¢/kWh by 2010 and 10¢/kWh by 2020. These potential reductions in cost, combined with the simplicity, versatility, reliability, and low environmental impact of PV systems, should help them to become increasingly important sources of economical premium-quality power over the next 20 to 30 years.

4.2 Solar-Thermal Systems

Solar-thermal power systems use various techniques to focus sunlight to heat an intermediary fluid, known as heat transfer fluid, which then is used to generate steam. The steam is used in a conventional steam turbine to generate electricity. At present, there are three solar-thermal power systems currently being developed: parabolic troughs, power towers, and dish/engine systems. Because these technologies involve a thermal intermediary, they can be readily hybridized with fossil fuels and in some cases adapted to utilize thermal storage. The primary advantage of hybridization and thermal storage is that the technologies can provide dispatchable power and operate during periods when solar energy is not available. Hybridization and thermal storage can enhance the economic value of the electricity produced and reduce its average cost.

Parabolic trough solar-thermal systems are commercially available. These systems use parabolic trough-shaped mirrors to focus sunlight on thermally efficient receiver tubes that contain a heat transfer fluid. This fluid is heated to about 390°C (734°F) and pumped through a series of heat exchangers to

produce superheated steam that powers a conventional turbine generator, producing electricity. Nine of these parabolic trough systems, built in the 1980s, are currently generating 354 MW in Southern California. These systems, sized between 14 and 80 MW, are hybridized with up to 25% natural gas in order to provide dispatchable power when solar energy is not available.

Power tower solar-thermal systems are in the demonstration and scale-up phase. They use a circular array of heliostats (large, individually tracking mirrors) to focus sunlight onto a central receiver mounted on top of a tower. The first power tower, Solar One, was built in Southern California and operated in the mid-1980s. This initial plant used a water/steam system to generate 10 MW of power. In 1992, a consortium of U.S. utilities joined together to retrofit Solar One to demonstrate a molten-salt receiver and thermal storage system. The addition of this thermal storage capability makes power towers unique among solar technologies by allowing dispatchable power to be provided at load factors of up to 65%. In this system, molten salt is pumped from a “cold” tank at 288°C (550°F) and then cycled through the receiver, where it is heated to 565°C (1049°F) and finally returned to a “hot” tank. The hot salt can then be used to generate electricity when needed. Current designs allow storage ranging from 3 to 13 hours.

5. HYDROPOWER

5.1 Introduction

Hydropower is currently the largest renewable resource used for electricity. It plays an essential role in many regions of the world, with more than 150 countries generating hydroelectric power. A survey in 1997 by *The International Journal on Hydropower & Dams* found that hydro supplies at least 50% of national electricity production in 63 countries and at least 90% in 23 countries. About 10 countries obtain essentially all their commercial electricity from hydro, including Norway, several African nations, Bhutan, and Paraguay.

There is about 700 GW of hydro capacity in operation worldwide, generating 2600 TWh/year (about 19% of the world’s electricity production). About half of this capacity and generation is in Europe and North America, with Europe the largest, at 32% of total hydro use, and North America at 23% of the total. However, this proportion is

declining as Asia and Latin America commission large amounts of new hydro capacity. Small, mini, and micro hydro plants (usually defined as plants generating less than 10 MW, 2 MW, and 100 kW, respectively) also play a key role in rural electrification in many countries. An estimated 300 million people in China, for example, depend on small hydro.

5.2 Capacity and Potential

There is vast unexploited potential worldwide for new hydro plants, particularly in the developing countries of Asia, Latin America, and Africa, but most of the best sites have already been developed in Europe and North America. There is also upgrading potential at existing schemes, though any future hydro projects will, in general, have to satisfy stricter requirements both environmentally and economically than they have in the past.

As shown in Table IV, the world's gross theoretical hydropower potential is about 40,000 TWh/year, of which about 14,000 TWh/year is technically feasible

for development and about 7000 TWh/year is currently economically feasible. The last figure fluctuates most being influenced not only by hydro technology, but also by the changing competitiveness of other energy/electricity options, the status of various laws, costs of imported energy/electricity, etc. The biggest growth in hydro generation is expected in the developing countries, where there is still a large potential for hydro development, but relatively little growth is expected in most OECD countries, where more than 65% of the economic potential is already in use. Thus, at least in the near future, hydro will likely remain the main source of electricity generation in developing countries that possess adequate water resources.

Until recent years, there has been less than 100 GW (about 350 TWh/year) of new hydro capacity under construction at any one time, equivalent to less than 15% of the capacity in operation. The figure has now risen, reflecting China's vast construction program, which includes the 18.2-GW Three Gorges Project, now in its second phase of construction. Most new hydro capacity is under

TABLE IV
Potentials of Hydroelectric Installed Capacities under Construction in 1997^a

Region	Gross theoretical potential (TWh/year)	Technically feasible potential (TWh/year)	Economically feasible potential (TWh/year)	Installed hydro capacity (GWe)	Hydro power production (TWh/year)	Hydro capacity under construction (MWe)
North America	5817	1510	912	141	697	882
Latin America and the Caribbean	7533	2868	1199	114	519	18,331
Sub-Saharan Africa	3294	1822	809	16	48	2464
Middle East and North Africa	195	216	128	9	27	7749
Western Europe	3258	1235	770	147	498	6707
Central and Eastern Europe	304	171	128	21	66	1211
Newly independent states of the former Soviet Union	3583	1992	1288	66	225	16,613
Centrally planned Asia and China	6511	2159	1302	64	226	51,672
South Asia	3635	948	103	28	105	13,003
Other Pacific areas	5520	814	142	14	41	4688
Pacific OECD countries ^c	1134	211	184	34	129	841
World total ^b	40,784	13,945	6964	655	2582	124,161

^a Source. WEA (2000).

^b Totals rounded.

^c OECD, Organization for Economic Cooperation and Development.

construction in Asia and South America. China has by far the most, with about 50 GW underway, and India, Brazil, and Russia each have more than 9 GW under construction. Brazil has the largest resources in world (800,000 GWh/year) of economically exploitable capacity and Norway depends almost entirely hydro for its electricity needs.

Hydropower continues to be the most energy-efficient way to generate electricity. Modern hydro turbines can convert as much as 90% of the available energy into electricity. The best fossil fuel plants are only about 50% efficient. In the United States, hydropower is produced for an average of 0.7¢/kWh. This is about one-third the cost of using fossil fuel or nuclear energy and one-sixth the cost of using natural gas. Hydro resources are also widely distributed compared with fossil and nuclear fuels and can help provide energy independence for countries without fossil fuel resources. There is also significant, widespread activity in developing small, mini, and micro hydro plants. At least 40 countries, particularly in Asia and Europe, have plants under construction and even more have plants planned. China, Brazil, Canada, Turkey, Italy, Japan, and Spain all have plans for more than 100 MW of new capacity.

5.3 Small Hydro

Small-scale hydro is mainly “run of river,” so does not involve the construction of large dams and reservoirs. It also has the capacity to make a more immediate impact on the replacement of fossil fuels because, unlike other sources of renewable energy, it can generally produce some electricity on demand (at least at times of the year when an adequate flow of water is available), with no need for storage or backup systems. It is also in many cases cost-competitive with fossil fuel power stations or (for remote rural areas) diesel-generated power.

Small hydro has a large, and as yet untapped, potential in many parts of the world. It depends largely on already proved and developed technology with scope for further development and optimization. Least-cost hydro is generally “high-head” hydro because the higher the head, the less the flow of water required for a given power level, and so smaller and less costly equipment is needed. Although this makes mountainous regions very attractive sites, they also tend to be in areas of low population density and thus low electricity demand, and long transmission distances often nullify the low-cost advantage. Low-head hydro, on the other hand, is relatively common, and also tends to be

found in or near concentrations of population, where there is a demand for electricity. Unfortunately, the economics also tend to be less attractive unless there are policy incentives in place to encourage their development.

5.4 Environmental and Social Impacts

Although hydroelectricity is generally considered a clean energy source, it is not totally devoid of greenhouse gas emissions and it can often have significant adverse socioeconomic impacts. There are arguments now that large-scale dams actually do not reduce overall GHG emissions when compared to fossil fuel power plants. To build a dam, significant amounts of land need to be flooded, often in densely inhabited rural area, involving large displacements of usually poor, indigenous peoples. Mitigating such social impacts represents a significant cost to the project, which, if it is even taken into consideration, often not done in the past, can make the project economically and socially unviable.

Environmental concerns are also quite significant, as past experience has shown. This includes reductions in biodiversity and fish populations; sedimentation, which can greatly reduce dam efficiency and destroy the river habitat; negative impacts on water quality; and contributions to the spread of water-related diseases. In fact, in the United States, several large power production dams are being decommissioned due to their negative environmental impacts. Properly addressing these issues would result in an enormous escalation of the overall costs for producing hydropower, making it far less competitive than is usually stated. As many countries move toward an open electricity market, this fact will come into play when decisions regarding investments in new energy sources are being made. If the large hydro industry is to survive, it needs to come to grips with its poor record of both cost estimation and project implementation.

5.5 Conclusions

Hydropower is a significant source of electricity worldwide and will likely continue to grow, especially in the developing countries. Although large dams have become much riskier investments, there still remains much unexploited potential for small hydro projects around the world. It is expected that growth of hydroelectricity will continue but at a slower rate than that of the 1970s and 1980s. Thus, the fraction of hydroelectricity in the portfolio of

primary sources of energy, which is today at 19%, is expected to decrease in the future. Improvements and efficiency measures are needed in dam structures, turbines, generators, substations, transmission lines, and environmental mitigation technology if the role of hydropower as a clean renewable energy source is to continue to be supported.

6. GEOTHERMAL ENERGY

6.1 Introduction

Geothermal energy, the natural heat within the earth, arises from the ancient heat remaining in Earth's core, from friction where continental plates slide beneath each other and from the decay of radioactive elements that occur naturally in small amounts in all rocks. The amount of geothermal energy is enormous. Scientists estimate that just 1% of the heat contained in the uppermost 10 km of Earth's crust is equivalent to 500 times the energy contained in all of Earth's oil and gas resources. Yet, despite the fact that this heat is present in practically inexhaustible quantities, it is unevenly distributed, seldom concentrated, and often at depths too great to be exploited industrially and economically.

Geothermal energy has been produced commercially for 70 years for both electricity generation and direct use; its use has increased rapidly during the past three decades. From 1975 to 1995, the growth rates worldwide for electricity generation and direct

use of geothermal energy were about 9%/year and about 6%/year, respectively. In 1997, geothermal resources had been identified in over 80 countries and there were quantified records of geothermal utilization in at least 46 countries.

6.2 Capacity and Potential

The worldwide use of geothermal energy amounts to about 44,000 GWh/year of electricity and 38,000 GWh/year for direct use, as shown in Table V. Although geothermal accounts for only about 0.3% of the total electrical power generated worldwide, among renewable sources (excluding hydro) it ranks first, at around 80% of the total. It is currently estimated that the geothermal potential of the world for electricity production is 12,000 TWh/year, and for direct use it is estimated to be even larger, at 600,000 EJ. A very small fraction of the geothermal potential has therefore been developed so far, and an accelerated use of geothermal energy in the near future is certainly feasible. In addition, the technology for both electricity generation and direct application is mature and accessible. Thus, the question of future geothermal development depends on whether it is economically competitive with other energy sources in the different markets around the world. Currently, the electric generation cost is variable, but averages around 4¢/kWh (U.S.); direct utilization production costs, though highly variable, are commonly under 2¢/kWh.

TABLE V

Geothermal Electricity Generation and Direct Use in 1997^a

Country	Electricity generation			Direct use		
	Installed capacity (MWe)	Total production		Installed capacity (MWt)	Total production	
		GWh/year	%		GWh/year	%
European Union	754	3,832		1031	3719	
Europe, other	112	471		4089	16,058	
Europe, total	866	4303	10	5120	19,777	52
North America	2849	16,249		1908	3984	
Central and South America	959	6869				
America, total	3808	23,118	53	1908	3984	10
Asia	2937	13,045	30	3075	12,225	32
Oceania	365	2901	6	264	1837	5
Africa	45	390	1	71	355	1
World total	8021	43,756		10,438	38,178	

^a Source. WEA (2000).

TABLE VI
Installed Geothermal Electricity Generation Capacity^a

Country	1990 MWe	1995 MWe	2000 MWe
Argentina	0.67	0.67	0
Australia	0	0.17	0.17
China	19.2	28.78	29.17
Costa Rica	0	55	142.5
El Salvador	95	105	161
Ethiopia	0	0	8.52
France (Guadeloupe)	4.2	4.2	4.2
Guatemala	0	33.4	33.4
Iceland	44.6	50	170
Indonesia	144.75	309.75	589.5
Italy	545	631.7	785
Japan	214.6	413.71	546.9
Kenya	45	45	45
Mexico	700	753	755
New Zealand	283.2	286	437
Nicaragua	35	70	70
Philippines	891	1227	1909
Portugal (The Azores)	3	5	16
Russia (Kamchatka)	11	11	23
Thailand	0.3	0.3	0.3
Turkey	20.6	20.4	20.4
United States	2774.6	2816.7	2228
Total	5831.72	6833.38	7974.06

^aData From the International Geothermal Association.

Table VI shows the installed geothermal electricity generation capacity in 1990, 1995, and 1998 for a selection of countries. The growth of the total generation capacity from 1990 to 1995 was about 16% and from 1995 to 1998, about 17%. The largest additions in generation capacity from 1990 to 1998 (in megawatts of electricity, MWe) were in the Philippines (957), Indonesia (445), Italy (224), Japan (315), Costa Rica (120), Iceland (95), United States (75), New Zealand (62), and Mexico (43). The most progressive of these countries, the Philippines, plans to add some 580 MWe to their installed capacity through 2008.

In the industrialized countries where installed electrical capacity is already very high, geothermal energy is unlikely in the near-term future to account for more than 1% of the total. On the other hand, in developing countries, with still relatively limited electrical consumption, but with good geothermal prospects, electrical generation from geothermal energy could make a significant contribution to the

TABLE VII
Top 10 Countries in Direct Utilization of Geothermal Energy^a

Country	Installed (MWt)	Production (GWh/year)
Japan	1159	7500
Iceland	1443	5878
China	1914	4717
United States	1905	3971
Hungary	750	3286
Turkey	635	2500
New Zealand	264	1837
France	309	1359
Italy	314	1026
Russia	210	673

^aSource. WEA (2000).

total installed electrical capacity. For example, 18% of the electricity in the Philippines already comes from geothermal sources; geothermal production is currently 17% in Nicaragua, 12% in El Salvador, 11% in Costa Rica, and 6% in Kenya.

Direct application of geothermal energy can involve a wide variety of end uses employing mostly existing technology. The technology, reliability, economics, and environmental acceptability of direct use of geothermal energy have been demonstrated throughout the world. In comparison with electricity production from geothermal energy, direct utilization has several advantages. It has higher energy efficiency (50–70% as opposed to 5–20% for conventional geothermal electric plants), generally a much shorter development time, and normally less capital investment is needed. In addition, direct application can use both high- and low-temperature geothermal resources and is therefore much more widely available in the world. Direct application is, however, much more site specific for the market, because steam and hot water are rarely transported long distances from the geothermal site. The longest geothermal hot water pipeline in the world, 63 km, is in Iceland. Space heating is the dominant (33%) type of direct geothermal use in the world; the various other types of direct use of geothermal energy include bathing/swimming/balneology (19%), greenhouses (14%), heat pumps (12%) for air cooling and heating, fish farming (11%), and industry (10%).

Table VII shows installed geothermal capacity and production in the top 10 direct-use countries in the world. It is worth noting that the two countries with the highest energy production, Japan and Iceland, are not the same as the two with the highest installed

capacities, China and the United States. The reason for this is the variety in the load factors for the different types of use.

6.3 Environmental Impacts

Geothermal fluids contain variable concentrations of gases, largely nitrogen and carbon dioxide, with some hydrogen sulfide and smaller proportions of ammonia, mercury, radon, and boron. Most of these chemicals are concentrated in the disposal water that is usually reinjected back into the drill holes, so that there is minimal release into the environment. The concentrations of the gases are usually low enough not to be harmful and the abatement of toxic gases can be managed with current technology. Carbon dioxide is the major component of the noncondensable gases in the steam, but its emission into the atmosphere per kilowatt-hour is well below the figures for power plants fired by natural gas, oil, or coal. Hydrogen sulfide is the pollutant of greatest concern in geothermal plants, yet even the sulfur emitted, with no controls, is only half of what is emitted from a coal-fired plant. Overall, with present technology able to control the environmental impact of geothermal energy development, it is considered to be a relatively benign source of energy.

6.4 Conclusions

Geothermal energy is generally cost-competitive with conventional energy sources and is produced by well-proved conventional technology. It is reliable and has been used for the more than half of the past century to heat large municipal districts, as well as to feed power plants generating hundreds of megawatts of electricity. It has strong potential to continue to expand, especially in the developing countries, and is a clean energy source that can help contribute to reducing our greenhouse gas emissions. It is felt that in the near term, the development of geothermal energy could help fulfill a bridging function during the next few decades, as other more modern clean fuel technologies and renewables mature enough to provide a meaningful share of the world energy supply.

7. RENEWABLE ENERGY SYSTEM COST AND PERFORMANCE

7.1 Recent Progress

There has been significant progress in cost reduction made by wind and PV systems, and biomass,

geothermal, and solar-thermal technologies are also experiencing cost reductions that are forecast to continue. Figure 8 presents forecasts made by the U.S. Department of Energy (DOE) for the capital costs of these technologies, from 1997 to 2030. Of course, capital costs are only one component of the total cost of generating electricity, which also includes fuel costs and operation and maintenance costs. In general, renewable energy systems are characterized by low or no fuel costs, although operation and maintenance costs can be considerable. It is important to note, however, that operation and maintenance costs for all new technologies are generally high, and can fall rapidly with increasing familiarity and operational experience. Renewable energy systems such as photovoltaics contain far fewer mechanically active parts than do comparable fossil fuel combustion systems, and therefore are likely in the long term to be less costly to maintain. Figure 9 presents U.S. DOE projections for the levelized costs of electricity production from these same renewable energy technologies, from 1997 to 2030. Given these likely capital and levelized system cost reductions, recent analyses have shown that additional generating capacity from wind and solar energy can be added at low incremental costs relative to additions of fossil fuel-based generation. These incremental costs would be further offset by environmental and human health benefits. Furthermore, a U.S. National Renewable Energy Laboratory (NREL) analysis shows that geothermal and wind energy could actually become more economic than coal in the next 15 years.

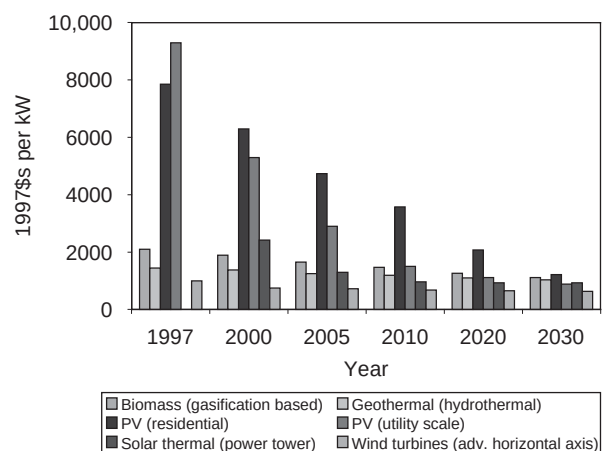


FIGURE 8 Capital cost forecasts for renewable energy technologies. PV, Photovoltaic. From the U.S. Department of Energy (1997).

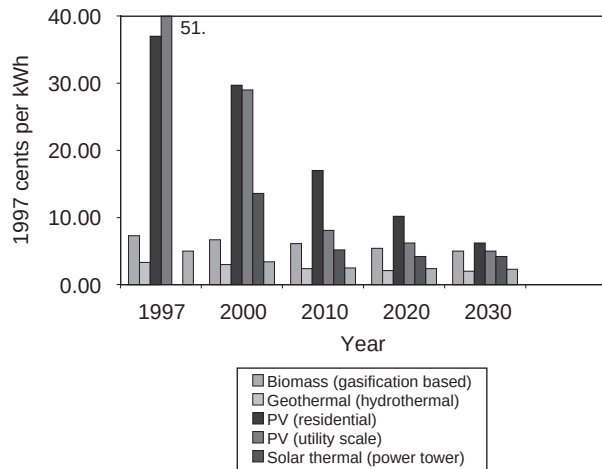


FIGURE 9 Levelized cost of electricity forecast for renewable energy technologies. PV, Photovoltaic. From the U.S. Department of Energy (1997).

Another analysis conducted by the Renewable Energy Policy Project (REPP) shows that adding 3050 MW of wind energy production in Texas, over a 10-year period, would entail only modest additional costs to residential customers. REPP estimates these additional costs to be about 75¢/month for a household using 1000 kWh/per month, or about \$9 annually. The economic case for renewables looks even better when environmental costs are considered along with capital and operating costs. As shown in Fig. 10, geothermal and wind can be competitive with modern combined-cycle power plants, and geothermal, wind, and biomass all have lower total costs compared to advanced coal-fired plants, once approximate environmental costs are also included. Shell Petroleum has made one of the highest profile projections of future renewables growth. As shown in Fig. 11, Shell projects that renewables could constitute about 15% of the OECD's energy production by 2020, and that renewables and natural gas combined could account for about 50% of total production.

As already noted, the remarkable difference between the setting for renewable energy today, relative to the past 30 years, is that renewable and other clean energy technologies are now becoming economically competitive, and the push to develop them is no longer being driven solely by environmental concerns. With regard to prospects for investing in companies developing clean energy resources, Merrill Lynch's Robin Batchelor recently stated that "This is not an ethical investment opportunity, it's a straightforward business opportunity." Batchelor also noted that the traditional energy

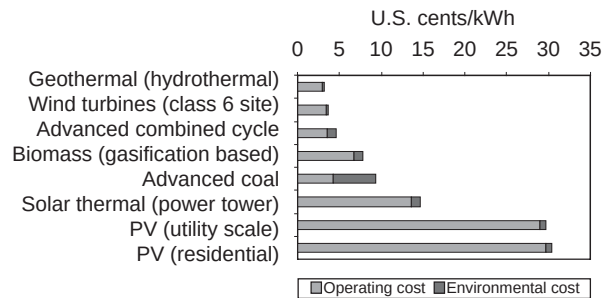


FIGURE 10 Actual electricity costs in 2000. PV, Photovoltaic. Data from the U.S. Department of Energy (1997, 2000).

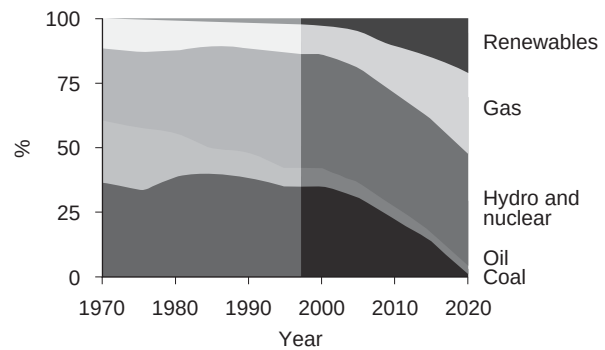


FIGURE 11 Electricity mix in Organization for Economic Cooperation and Development countries. From Shell Petroleum (2000).

sector has lacked appeal to investors in recent years because of heavy regulation, low growth, and a tendency to be highly cyclical. He has identified 300 companies worldwide whose aim is to develop wind, solar, and wave power technologies and to advance capabilities in energy storage, conservation, and on-site power generation. Further evidence of the impending transition to renewable energy systems can be found in recent corporate reorganization among the world's largest oil companies. Corporate giants such as Shell and BP/Amoco, which would now like to be known as "energy companies" rather than "oil companies," have recently reorganized into a broader array of business units, including those exclusively focused on renewables. Such subsidiary units now include "Shell Renewables," and "BP Solar."

7.2 Lessons Learned in Developing Countries

In developing nations, renewable energy technologies are increasingly used to address energy shortages and to expand the range of services in both rural and urban areas. In Kenya, over 120,000 small

(20–100 W_p) (W_p, photovoltaic-derived watts) solar-PV systems have been commercially financed and installed in homes, battery charging stations, and other small enterprises. Meanwhile, a government program in Mexico has disseminated over 40,000 such systems. In the Inner Mongolian autonomous region of China, over 130,000 portable windmills provide electricity to about one-third of the non-grid-connected households in this region. These examples demonstrate that the combination of sound national and international policies and genuinely competitive markets—the so-called level playing field—can be used to generate sustainable markets for clean energy systems. They also demonstrate that renewable energy systems can penetrate markets in the developing world, even where resources are scarce, and that growth in the renewables sector need not be limited to applications in the developed world. Just as some developing countries are bypassing construction of telephone wires by leaping directly to cellular-based systems, so too might they avoid building large, centralized power plants and instead develop decentralized systems. In addition, to help mitigating the environmental costs of electrification, this strategy can also reduce the need for the construction of large power grids.

In the absence of policy interventions, we may remain locked-in to existing technologies, even if the benefits of technology switching overwhelm the costs. There are numerous examples, however, of an entrenched or locked-in technology being first challenged and ultimately replaced by a competing technology. This process is generally enabled by a new wave of technology, and it is sometimes achieved through a process of hybridization of the old and the new. Technological “leapfrogging” is another possibility, but this seems to occur relatively rarely. A prime example of the hybridization concept is in the case of the competition between gas- and steam-powered generators, dating back to the beginning of the century. From about 1910 to 1980, the success of steam turbines led to a case of technological lock-in, and to the virtual abandonment of gas turbine research and development. However, partly with the aid of “spillover” effects from the use of gas turbines in aviation, the gas turbine was able to escape the lock-in to steam turbine technology. First, gas turbines were used as auxiliary devices to improve steam turbine performance, and then they slowly became the main component of a hybridized, “combined-cycle” system. In recent years, orders for thermal power stations based primarily on gas turbines have

increased to more than 50% of the world market, up from just 15–18% in 1985.

7.3 Leveling the Playing Field

As was shown in Fig. 10, renewable energy technologies tend to be characterized by low environmental costs. In an ideal world, this would aid them in competing with conventional technologies, but of course many of these environmental costs are “externalities” that are not priced in the market. Only in certain areas and for certain pollutants do these environmental costs enter the picture, and clearly further internalizing these costs would benefit the spread of renewables. The international effort to limit the growth of greenhouse emissions through the Kyoto Protocol may lead to some form of carbon-based tax, and this could prove to be an enormous boon to renewable energy industries. However, support for the Kyoto Protocol among industrialized countries remains relatively weak, particularly in the United States, and any proposed carbon-based taxation scheme will surely face stiff opposition.

Perhaps more likely, concern about particulate matter emission and formation from fossil fuel power plants will lead to expensive mitigation efforts, and this would help to tip the balance toward cleaner renewable systems. In a relatively controversial move, the U.S. Environmental Protection Agency (EPA) has recently proposed new ozone and particulate matter (PM) standards that are even more stringent than the current standards that remain unattained in some U.S. urban areas. The EPA has justified these new regulations with analysis that shows that new standards are necessary to provide increased protection against a wide range of potential health impacts. For example, the EPA estimates that even if Los Angeles County (California) were to meet the existing PM standards, 400 to 1000 deaths per year would still occur as a result of exposure to very fine PM (under 2.5 μm in diameter), which presently is not regulated. The combination of increased pressure to attain ozone and PM standards will further complicate siting of new fossil-fueled power plants in some areas of the United States, particularly where it is now required to find “offsets” for pollution impacts. This will indirectly but surely benefit renewable energy technologies, which do not typically face these difficulties in obtaining siting permits.

7.3.1 Public and Private Sector Investment Issues

A fundamental problem with any new technology is that, by definition, it does not have the track record

of performance that exists for older, more established systems. Proponents of existing technologies in mistaken arguments against technological change often cite this fact. New technologies and operational procedures do present greater risks, but at the same time present greater opportunities for innovation and profit. Emerging energy systems were long seen as an area of risky investments, with the history of renewable energy systems seen as the primary illustration of that “fact.” It has been argued that this pattern is not only illusory, but is largely a self-fulfilling prophecy. Larger gains, both to individual companies and to society, typically stem from carefully targeted but consistently pursued avenues of research, innovation, and implementation. Renewable energy systems offer this same combination of increased uncertainty, great promise, and the potential for significant innovations and profits.

There are two principal rationales for government support of research and development to develop clean energy technologies (CETs). First, conventional energy prices generally do not reflect the social cost of pollution. This provides a rationale based on a well-accepted economic argument, for subsidizing research and development for CETs as potential alternatives to polluting fossil fuels. Second, private firms are generally unable to appropriate all the benefits of their research and development investments. Consequently, the social rate of return for research and development exceeds available private returns, and firms therefore do not invest enough in research and development to maximize social welfare. Thus, innovation “spillover” among CET firms is a form of positive externality that justifies public research and development investment. The conventional wisdom is that government should restrict its support to research and development and let the private sector commercialize new technologies. Failed CET commercialization subsidies (e.g., the U.S. corn ethanol and synfuels programs) bolster this view. Nonetheless, there are compelling arguments for public funding of market transformation programs (MTPs) that subsidize demand for some CETs in order to help commercialize them. Further, the argument that it may not be worthwhile for firms to invest in new technologies because of the spillover effects is generally false as well. Early investment in new technologies in promising market sectors has proved to be the best strategy for firms interested in long-term (and not simply short-term) profitability.

A principal motivation for considering MTPs is inherent in the production process. When a new technology is first introduced, it is invariably more

expensive than established substitutes. There is, however, a clear tendency for the unit cost of manufactured goods to fall as a function of cumulative production experience. Cost reductions are typically very rapid at first, but taper off as the industry matures. This relationship is called an “experience curve” when it accounts for all production costs, and it can be described by a progress ratio (PR), where unit costs fall by $100 \times (1 - \text{PR})$ percent with every doubling of cumulative production. Typical PR values range from 0.7 to 0.9 and are widely applicable to technologies such as toasters, microwave ovens, solar panels, windmills, and essentially any good that can be manufactured in quantity. For example, Fig. 12 presents PRs for photovoltaics, windmills, and gas turbines. All three have initial PRs of approximately 0.8, which is a typical value observed for many products. After 1963, the gas turbine PR increased substantially, however, indicating an attenuation of experience effects.

The benefits of production experience may accrue primarily to the producing firm, or they may spill over to its competitors. Among other channels, “experience spillovers” could result from hiring competitors’ employees, reverse engineering rivals’ products, informal contacts among employees of rival firms, or even industrial espionage. If firms retain the benefits of their own production experience, they have an incentive to consider experience effects when deciding how much to produce. Consequently, they will “forward price,” producing at a loss initially to bring down their costs and thereby maximize profit over the entire production period. It has been shown that, in each period, firms maximize profit by setting marginal revenue equal to their true marginal cost (TMC). For a firm with a discount rate of zero, TMC is defined as the cost of the last unit the firm plans to produce. If the firm has a positive discount rate, its TMC will be somewhat higher. However, due to experience effects, the TMC is always lower than current marginal cost (except

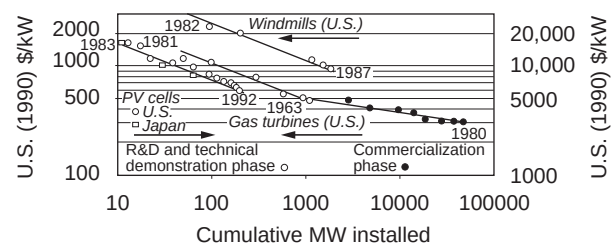


FIGURE 12 Progress ratios for photovoltaics (PV), windmills, and gas turbines. R&D, Research and development. From IIASA/WEC (1995).

for the very last unit produced). In addition, unless there is 100% spillover, the first-best social optimum requires regulating a single firm such that it produces in every period up to the level where price equals true marginal cost. This policy would require subsidizing the monopolist because it would be selling units for less than their current marginal cost in every period except the last. Absent regulation, when experience spillover is low, cumulative production experience gives incumbent firms widening cost advantages over potential entrants. Thus, early entrants can cut back their output and raise prices above the competitive level without prompting other firms to enter. The incentive to forward price partially mitigates the problem, but output remains below the social optimum. When experience does spill over substantially, firms still do not produce enough to maximize social efficiency. This follows because experience becomes a positive externality, analogous to the research and development appropriability problem. Thus, firms insufficiently forward price because they do not value the portion of their experience benefits that accrue to other firms.

To summarize, regardless of the level of spillover, strong experience effects imply that output is less than the socially efficient level. That is, some consumers would be willing to pay more than the cost of additional production, but the market fails to facilitate these mutually beneficial trades. Market transformation programs can improve social welfare by correcting the output shortfall associated with these experience effects. Moreover, as with research and development, MTPs also help to promote the use of clean energy technologies as alternatives to polluting fossil fuel technologies, and thereby reduce the social costs of pollution. When politically possible, the first-best policy is to internalize pollution costs fully (e.g., through pollution taxes set at the marginal social cost of the pollution externality or through tradable emissions permits set at the socially optimal pollution level). Governments chronically fail to achieve this, however, providing another clear rationale to support market transformation programs.

When evaluating MTPs, it is essential to account for positive feedback between the demand response and experience effects. An MTP increases the quantity produced in the first year and, due to experience effects, year 2 unit costs are lower than they would have been without the additional production from the MTP. These lower costs, in turn, imply that the quantity demanded in year 2 is

higher. This “indirect demand effect,” in turn, adds to cumulative production experience and further lowers unit costs in future years. This process continues indefinitely, though it gradually dissipates once the MTP is discontinued. Accounting for these indirect demand effects substantially raises the benefit–cost ratio (BCR) of typical MTPs. Even without accounting for environmental benefits, the case studies of MTPs targeting photovoltaics and efficient lighting show positive BCRs of 1.05 and 1.54, respectively.

These results suggest a role for MTPs in national and international technology policies; however, the costs of poor program design, inefficient implementation, or simply choosing the “wrong” technologies can easily outweigh cost reduction benefits. This suggests that MTPs be limited to emergent CETs with a steep industry experience curve, a high probability of major long-term market penetration once subsidies are removed, and a price elasticity of demand of approximately unity or greater. The condition that they be clean technologies mitigates the risk of poor MTP performance by adding the value of displaced environmental externalities. The other conditions ensure a strong indirect demand effect. Finally, as with energy research and development policy, public agencies should invest in a portfolio of new clean energy technologies in order to reduce overall market transformation program performance risk through diversification.

8. CONCLUSIONS

In conclusion, it is clear that the promise of renewable energy has now become a reality. Both solar photovoltaics and wind energy are experiencing rapid sales growth, declining capital costs and costs of electricity generated, and continued performance increases. Because of these developments, market opportunity exists now both to innovate and to take advantage of emerging markets, with the additional assistance of governmental and popular sentiment. The development and use of these sources can enhance diversity in energy supply markets, contribute to securing long-term sustainable energy supplies, make a contribution to the reduction of local and global atmospheric emissions, provide commercially attractive options to meet specific needs for energy services (particularly in developing countries and rural areas), and create new employment opportunities.

Acknowledgments

It is a pleasure to thank the Energy Foundation for support, as well as the University of California Energy Institute. This work draws on previous work done by the author in conjunction with two of his past postdoctoral fellows, Dr. Antonia Herzog, now of the Natural Resources Defense Council, and Dr. Timothy Lipman, now a research fellow at both the Berkeley and the Davis campuses of the University of California.

SEE ALSO THE FOLLOWING ARTICLES

Biodiesel Fuels • Biomass for Renewable Energy and Fuels • Computer Modeling of Renewable Power Systems • Energy Efficiency, Taxonomic Overview • Public Reaction to Renewable Energy Sources and Systems • Renewable Energy and the City • Renewable Energy in Europe • Renewable Energy in Southern Africa • Renewable Energy in the United States • Renewable Energy Policies and Barriers • Renewable Portfolio Standard

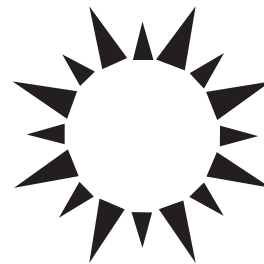
Further Reading

- American Wind Energy Association. (2003). "Global Wind Energy Market Report." Available on the Internet at <http://www.awea.org>.
- American Wind Energy Association. (2003). "Wind Energy Fact Sheet: Comparative Costs of Wind and Other Fuels." Available on the Internet at <http://www.awea.org>.
- Duke, R. D., and Kammen, D. M. (1999). The economics of energy market transformation initiatives. *Energy J.* 20(4), 15–64.
- Ezzati, M., and Kammen, D. (2001). Indoor air pollution from biomass combustion and acute respiratory infections in Kenya: An Exposure–response study. *Lancet* 358, 619–624.
- Hall, D. O., and Scrase, J. I. (1998). Will biomass be the environmentally friendly fuel of the future? *Biomass Bioenergy* 15(4/5), 357–367.
- Intergovernmental Panel on Climate Change (IPCC). (1996). "Climate Change 1995: Scientific–Technical Analyses of Impacts, Adaptations, and Mitigation of Climate Change" (Watson, Zinyowera, and Moss, Eds.), Chap. 19. Cambridge University Press, Cambridge.
- International Institute for Applied Systems Analysis/World Energy Council (IIASA/WEC). (1995). "Global Energy Perspectives to 2050 and Beyond" (N. Nakicenovic and M. Jefferson, Eds.). IIASA/WEC, Laxenburg, Austria and London.
- Johansson, T. B., Kelly, H., Reddy, A. K. N., and Williams, R. H. (eds.). (1993). "Renewable Energy: sources for Fuels and Electricity." Island Press, Washington, D.C.
- Kammen, D. M., Bailis, R., and Herzog, A. V. (2001). "Clean Energy for Development and Economic Growth: Biomass and Other Renewable Energy Options to Meet Energy and Development Needs in Poor Nations." United Nations Development Programme (UNDP), Report for the 7th Conference of the Parties to the UN Framework Convention on Climate Change (COP7-UNFCCC). Marrakech, Morocco, October 29–November 9. UNDP, New York.
- Margolis, R., and Kammen, D. M. (1999). Underinvestment: the energy technology and R&D policy challenge. *Science* 285, 690–692.
- Maycock, P. (1998). *PV News*.
- Maycock, P. (2000). The world PV market 2000: Shifting from subsidy to 'fully economic'? *Renewable Energy World* 3(4), 59–73.
- Overend, R. P. (2000). National Renewable Energy Laboratory, Washington, D.C.
- President's Committee of Advisors on Science and Technology (PCAST) Energy R&D Panel. (1997). "Federal Energy Research and Development for the Challenges of the Twenty-First Century." Report of the Energy Research and Development Panel of the President's Committee of Advisors on Science and Technology, November. U.S. Dept. of Energy, Washington, D.C.
- Shell Petroleum. (2000). "Global Scenarios: 1998–2020." Summary brochure. Group Planning, Shell International, London.
- Sivan, K., and Larson, E. D. (2000). "Bioenergy Primer: Modernised Biomass Energy for Sustainable Development." United Nations Development Programme (UNDP), Energy and Atmosphere Programme. UNDP, New York.
- U.S. Department of Energy (DOE). (1997). "Renewable Energy Technology Characterizations." Topical report, prepared by U.S. DOE Office of Utility Technologies and EPRI, TR-109496, December. U.S. Dept. of Energy, Washington, D.C.
- U.S. Department of Energy (DOE). (2000). "Annual Energy Outlook 2000." DOE/EIA-0383(00), December. Energy Information Administration, Washington, D.C.
- World Energy Assessment (WEA). (2000). "Energy and the Challenge of Sustainability," Chaps. 5 and 7. Report, prepared by United Nations Development Programme, United Nations Department of Economic and Social Affairs and World Energy Council, September. United Nations Publications, New York.
- Worldwatch Institute. (1999). "Vital Signs 1999. The Environmental Trends That Are Shaping Our Future." W. W. Norton, London.

Renewable Portfolio Standard

MARK JACCARD

Simon Fraser University
Vancouver, British Columbia, Canada



-
1. Rationale
 2. Design Considerations
 3. Early Adopters
 4. Preliminary Assessment

Glossary

fixed feed-in tariff An offer by government or a utility to purchase electricity from renewables producers at a fixed price, regardless of the producers' costs of production.

green (renewables) certificate A certificate, provided by a regulator to renewables producers for specified units of their electricity output; can be traded to other electricity producers who need it to meet their obligation to produce a minimum percentage of renewable electricity under the renewable portfolio standard.

quantity-based policy A policy that specifies a quantity outcome (as opposed to a price-based policy such as an environmental tax or subsidy); an example is the renewable portfolio standard policy, which requires production of a specific quantity of electricity.

renewable energy Forms of energy (sunlight, wind, geothermal heat, wave power, tidal energy, hydropower, and organic matter) that flow through Earth's biosphere, available for human use indefinitely, provided that the physical basis for their flow is not destroyed.

The renewable portfolio standard (RPS) is a policy instrument to force greater production of electricity from energy sources (usually renewable) that have desirable social and environmental characteristics. By requiring market participants to deliver a minimum amount of electricity, typically as a percentage of total sales, from specific fuels and/or technologies, the RPS is a "quantity-based policy." In Europe, the policy is often referred to as "green certificate" trading. Renewable energy sources are normally defined as sunlight, wind, geothermal heat, wave

power, tidal energy, hydropower, and organic matter, but the definition used in a given jurisdiction's RPS may be narrower or broader, depending on local views of the social and environmental characteristics of each technology and fuel.

1. RATIONALE

Renewables appear to have lower social and environmental impacts and risks when compared with electricity derived from conventional sources, such as fossil fuels and nuclear power. Environmental benefits may include reduced emissions contributing to smog or producing acid rain and greenhouse gases relative to a total reliance on fossil fuels, and zero risk of radiation exposure or nuclear weapons proliferation relative to nuclear power. Greater use of renewable sources may also enhance energy security by increasing supply diversity and the use of indigenous fuels, thereby reducing vulnerability to supply disruptions and volatile fuel prices. Renewables can also support economic development through reducing payments for imported fuels (increasing local multiplier effects) and fostering the development of local technical expertise.

Renewables are usually more expensive than conventional electricity sources when compared on a financial cost basis. Because of this, both monopoly and competitive electricity producers have concentrated their investment on conventional electricity technologies, with renewables usually accounting for only a small percentage of the generating stock. The only countries dependent on renewables for electricity generation are those with substantial, low-cost hydropower endowments, such as Brazil, Norway, New Zealand, and Canada. But because many large hydropower facilities disrupt ecosystems, sometimes dramatically, hydro plants exceeding a threshold of, say, 20 MW may be excluded from RPS eligibility.

There are three arguments for why the financial comparison of renewables and conventional generation sources does not reflect the full costs and benefits to society of the alternative sources:

1. Some jurisdictions provide greater subsidies to conventional generation fuels and technologies.
2. The full costs of pollution (externalities) are not included in the financial costs of conventional electricity sources, and are usually much greater than pollution costs for renewables.
3. Although humans have used some renewables (such as wind) for millennia, renewables are today associated with new technologies, the relative costs of which will fall with widespread commercialization because of economies of learning and economies of scale in manufacture.

Thus, although from a strict financial cost perspective (Fig. 1, left side) most renewables are more expensive compared to conventional electricity, the situation may be reversed when these three additional factors are taken into account (Fig. 1, right side). Inclusion of these three factors is referred to as the “long-run, social cost perspective.” Whereas it is not certain that renewables are always less expensive on a social cost basis, they are nonetheless recognized as desirable sources of electricity that will have difficulty penetrating the market given the pressure on electricity investors to focus on financial considerations. This constraint is exacerbated if electricity markets continue the current trend toward independent electricity production, mostly by privately owned companies.

In recognition of the possible advantages of renewables on a social cost basis, governments have developed policies to improve the financial competitiveness of renewables. There are four broad categories of measures to support renewables:

1. Assist the commercialization of renewables with indirect support. This may include funding of

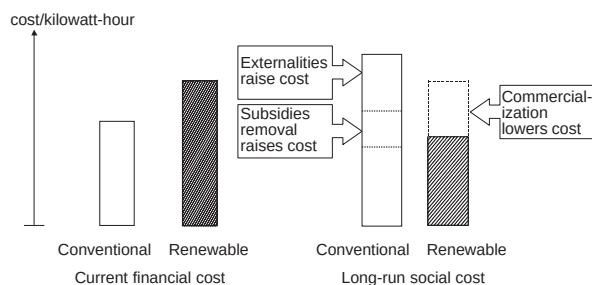


FIGURE 1 Conventional versus renewable electricity: financial and social cost comparison.

demonstration projects, audits, evaluations, resource assessments, research and development support, and training.

2. Provide direct financial support to renewables. This may include capital grants, preferential purchase prices, tax advantages, or low-interest loans. Energy price subsidies can be in the form of preferential fixed-price tariffs (feed-in tariffs), a fixed premium on the prevailing electricity price, or a premium payment to renewable electricity producers determined by a “competitive bidding” process.

3. Increase the cost of polluting sources of supply. This may include the reduction or elimination of subsidies to conventional sources, higher energy or pollution taxes (externality taxes), stronger technology regulations, or emission cap regulations.

4. Foster voluntary or mandatory minimum market shares for renewables. This may include voluntary agreements with producers, green tariffs that allow consumers to pay extra for renewables, and the RPS.

Since the energy crisis of the 1970s, governments have especially relied on the first two mechanisms to support renewables. Various indirect support policies (e.g., research and development subsidies) have been in place throughout the past 30 years. But the dominant policy in terms of impact has been direct financial support. Many governments have provided tax credits for renewables. Electric utilities and governments have offered above-market, fixed-tariff prices to acquire electricity from renewables producers. In the United States, the Public Utility Regulatory Policy Act of 1978 required utilities to purchase renewables and other independent electricity at prices that reflected the long-run cost of new, high-cost nuclear and fossil fuel plants. In Europe, fixed-tariff price support was implemented through direct government subsidy (e.g., in Denmark) or cross-subsidy from other utility customers (e.g., in Germany). In the 1990s, however, interest in the third and fourth approaches has increased as governments have explored policies that raise the cost of polluting sources of supply. Europe has focused on pollution regulations and externality taxes, and the United States has focused on regulations and systems that involve emissions caps and tradable permits. Industrialized countries, and increasingly developing countries (e.g., China), have also reduced subsidies to conventional fossil fuel and nuclear energy producers. Voluntary green tariffs and the RPS

emerged in the late 1990s in Europe, the United States, and Australia.

Three characteristics help explain the emerging interest in the RPS: First, an RPS maintains continuous incentives for renewable producers to seek cost reductions (economic efficiency) and can be designed to ensure that these cost reductions are passed on to consumers (equity). This is achieved by mechanisms that establish continuous cost competition among renewable producers for their share of the RPS. Second, because an RPS ensures the attainment of a specific market share, it can be directly linked to government policy objectives. Governments can set an RPS target as part of a package of policies to meet an environmental target such as CO₂ reduction. Third, an RPS minimizes government involvement relative to other measures. The government's budget is not implicated because customers pay producers directly the extra financial cost of renewables, and the selection of winning bids can be left to market forces (competitive bidding) instead of government evaluation. Because the RPS has only recently emerged as a policy tool, there is insufficient experience to provide an empirical evaluation of its performance relative to key alternatives, such as subsidized feed-in tariffs and competitive bidding for subsidies. However, critical issues in its design have already emerged.

2. DESIGN CONSIDERATIONS

Many considerations influence the design of an RPS. Major categories include selection of the target, determining resource eligibility, determining where and to whom the policy is applied, designing flexibility mechanisms, and setting administrative responsibilities. Although these are presented individually here, they cannot be considered in isolation from each other in the design process. A clear set of policy objectives is required for navigating these considerations. These include standard policy objectives, such as effectiveness, economic efficiency, equity, political acceptability, and administrative feasibility, but also the specific environmental and perhaps social goals that are the motivation for the RPS in the first place. For example, the RPS has been linked to the greenhouse gas reduction goals of several jurisdictions. The following discussion of design considerations assumes that the policy objectives include the standard list, as well as those environmental objectives that tend to favor renewables relative to fossil fuel- and nuclear-generated electricity.

2.1 Target Selection

The main challenge of the RPS is determination of the binding target or quota for renewable energy production. There are several aspects to this determination.

2.1.1 Target Size

Setting target size requires consideration of impacts on different objectives, including anticipated costs. These impacts will depend in part on the local cost and availability of renewables and the price of conventional electricity sources. From a political acceptability perspective, the target would be large enough to move the industry toward the environmental objective, but not so large that it results in a dramatic increase in electricity prices. The size of the target can also affect the level of competition among fuels or technologies. For example, if targets could be achieved with a relatively small number of large projects, there may be less competition and little pressure to innovate. Some knowledge of the potential costs and quantities of renewables is therefore important in establishing targets.

2.1.2 Target Timing

Another consideration in selecting the target is timing. Some lead time may be required to permit cost-effective responses to the RPS target. The length of the lead time will depend on the magnitude of the target and the characteristics of eligible resources. If existing facilities are not eligible for meeting targets, sufficient lead time will be required for investments in new facilities. A phase-in period would enable the government to conduct interim evaluations in order to assess impacts and make target adjustments where necessary, thereby maintaining fairly stable compliance costs. However, duration and stability of targets can also be important so that buyers and sellers can more easily negotiate long-term supply contracts. Long-term commitments that guarantee a revenue stream are important for securing low-cost project financing, especially for capital-intensive technologies such as renewables.

2.1.3 One or Multiple Renewables Targets

An issue in target setting is whether there should be one target for which all renewables compete, or a separate target for different classes of renewables. Under the first approach, the lowest cost options would be developed to meet the aggregate target. This approach should advance environmental goals more cost-effectively, but it might do less well in

furthering other objectives, such as greater supply diversity or increased support for newly emerging, but still expensive, types of renewables. Perhaps one renewable (e.g., wind) will capture the entire RPS market even though other renewables would eventually become competitive were they assisted in achieving commercialization and production thresholds. The state of Connecticut designed its RPS with separate targets for Class 1 renewables (solar, wind, landfill gas, and sustainable biomass) and Class 2 renewables (hydropower and municipal solid waste), the intent being to limit the share of hydropower and municipal solid waste within the RPS.

2.1.4 Target Adjustment and Other Cost Cap Measures

If the RPS is set too high, it could substantially increase the price of electricity, with perhaps significant political acceptability and economic performance impacts. There are several ways in which the government (or RPS administrator) could deal with this challenge. First, the RPS target could be voluntary (as in The Netherlands in its initial policy), or converted to voluntary status as higher-than-anticipated costs emerge. Second, the government could adjust the target downward as evidence of high cost emerged. Third, the government could set a modest penalty that participants would pay to the extent that they failed to comply completely with the RPS. This penalty provides assurance to participants that costs will not rise above the penalty level. Fourth, if the RPS includes a system of tradable renewable certificates (green certificates in Europe), the government could offer to sell an unlimited number of these at a modest price, in effect providing a price cap (e.g., The Netherlands). These measures all ensure that costs will not exceed a given level, but at the possible expense of not achieving the RPS target, and the corresponding environmental target associated with it. On the negative side, adjusting targets could jeopardize the credibility and predictability of the program, with serious consequences for its long-term success.

2.2 Eligible Resources

In addition to specific targets, the RPS must establish resource eligibility. This will depend on the objectives for the RPS and the local viability of different types of resources.

2.2.1 Renewables vs Other Desired Technologies

A key eligibility decision is whether the RPS should be limited to renewable sources of electricity or can

include other energy/technology combinations. This depends on the objectives and, in some sense, on the comfort level with using the term “renewables” to symbolize the broader concept of environmentally desirable technologies. In Italy, only truly renewable forms of energy are included in the RPS definition. In some jurisdictions, however, fossil fuel-driven cogeneration or fuel cell technology is included because of the higher energy efficiency and lower local emissions of these mechanisms. In proposing its RPS, Australia included solar water heaters that displace electric water heaters that use electricity from fossil fuel plants.

2.2.2 Existing Renewables vs New Investments

Another eligibility consideration is whether the RPS should apply to all renewable resources or only to new investments in renewable resources. Applying it to new investment helps advance the new technology commercialization intent of the RPS, as adopted in the Italian RPS. However, this may increase administrative complexity by requiring ongoing monitoring and regulatory rulings as to whether a given expenditure is just maintenance and refurbishment of an existing facility or a truly new investment (a problem experienced in a different context with the U.S. Clean Air Act). Is a complete reconstruction of a small hydropower facility a new or old investment? If new, what if only the turbines are replaced? If, however, the policy goal is simply to expand output from renewables at the lowest cost, the government may be indifferent regarding investments for reconditioning or for totally new facilities.

2.2.3 Grid-Connected Renewables vs All Renewables

The RPS may apply only to grid-connected facilities or to both grid-connected and off-grid facilities. Although the simplest approach is to focus the policy on changing the grid-transmitted market share of renewables, some would argue that environmental objectives are equally served by the development of renewables anywhere, including off grid. However, administrative realities must be considered. It would be difficult to verify the annual production of a myriad of small off-grid generators, leading to high administrative costs.

2.2.4 Facility Size

The size of resources may also be part of the RPS eligibility determination. Size limits can affect the number of potential projects required to meet the

overall target. They can also indirectly affect objectives such as diversity and the location of projects. The RPS of the state of Maine in the United States had an upper limit of 100 MW, which applied to both existing and new facilities. A limit of 10 or 20 MW is sometimes suggested as the maximum size that should be allowed for hydropower projects to be included in the RPS.

2.2.5 Imports vs Domestic Renewables

External, renewable electricity producers may wish to sell into a jurisdiction with an RPS standard. Again, determination of eligibility depends on the policy objectives. If the expansion of domestic resources is key, imports of renewables may be limited or excluded from the RPS, which does not prevent them from competing in the non-RPS market against conventional generation facilities. In many jurisdictions, external participation in the RPS will require some consideration of existing trade agreements and commercial law, which may limit policymakers' ability to discriminate between domestic supplies and imports. Both Europe and North America are working toward electricity trade rules that should make external participation easier.

2.3 Applicability

The applicability of the RPS target is another important design consideration. Key issues include the geographic coverage of the standard, the applicability of the standard to specific market participants, and whether it should be applied to capacity or energy.

2.3.1 Geographic Coverage

In terms of geographic coverage, the RPS may be most effective if the breadth of its application is matched to the breadth of environmental and other objectives. This explains why the European Union (EU), which sees renewables as a component of its greenhouse gas emission reduction objective, is seeking to achieve an EU-wide RPS. However, the physical limitations to trade must be considered. In some countries, the lack of grid interconnection makes implementation of even a national RPS problematic, unless some form of credit trading is permitted. In these cases, it may be easier initially to set an RPS for each major grid area.

2.3.2 Market Participants Regulated by the RPS

The RPS must specify the market participants to which it applies. Possible participants include

producers, wholesale customers or brokers, retail suppliers, distribution utilities, and consumers. Some jurisdictions want the RPS to be associated with final consumption, hoping that the public will better understand its link to environmental or other policy objectives. Denmark's proposed RPS applies to consumers, although distribution utilities or retail suppliers would presumably represent the interests of small consumers. Others believe that the administrative ease of regulating producers is important. Italy has size restrictions so that the standard applies to only about 100 producers nationwide.

2.3.3 Energy Production vs Installed Capacity

The RPS may apply to energy production or installed capacity. Capacity may be easier to verify, but energy production is directly connected to environmental objectives. An energy-determined RPS provides the incentive for producers to maximize production from individual projects, and thereby lower costs.

2.4 Flexibility Mechanisms

An RPS can be less costly to implement, yet still achieve the same target, if there is flexibility in how it is applied both to individual producers and across the entire market.

2.4.1 Account-Balancing Mechanisms for Individual Producers

One way to increase flexibility is to allow those to whom the policy is applied to have some flexibility in the choice of calendar year and in the period for achieving their annual RPS. This is known as an "account-balancing mechanism." For example, the inherent output variability of some types of renewables, such as wind, hinders accurate fulfillment of the RPS target each year. The electricity provider can be given extra months to meet the previous year's commitment as long as the full RPS requirement is achieved on average over a period of years. The Australian RPS has a 3-year balancing mechanism for electricity providers.

2.4.2 Trading Mechanisms between Producers

Another way to increase flexibility is to allow trade among electricity providers so that the RPS target is met in aggregate, even if some providers exceed the RPS and others fail to achieve it. This approach can significantly decrease the total cost of the

RPS because it allows the maximum electricity production from those renewables generators with the lowest costs. This trading can be facilitated by the creation of tradable certificates, hence the European term “green certificate” trading for the RPS. As of 2003, The Netherlands, Italy, England, Australia, and some U.S. states (Texas) had such certificates. The creation of certificates is not essential, however, for trading to occur. The RPS administrator could allow electricity providers to use a variety of instruments to demonstrate achievement of the RPS target in aggregate while they trade among themselves. Contracts between providers could be filed with the administrator and these would be verified against actual inputs to the grid.

2.5 Administrative Responsibilities

There are several administrative responsibilities in implementing a RPS. These could all be handled by one RPS administrator or delegated to specialized agencies.

2.5.1 *Setting the RPS Target*

The RPS target would likely be set at a senior level of government because it involves critical trade-offs between environmental targets, economic development goals, and concerns about electricity prices.

2.5.2 *Certification of Renewables*

Certification of renewables could be delegated to the operator of the grid, subject to periodic review by the government agency responsible for administering the RPS. Environmental regulators could help with the definition of what is a renewable resource, although, again, this may be a high-level policy decision.

2.5.3 *Compliance Monitoring*

Compliance monitoring could be delegated to the operator of the grid, because this entity will already have information on production of connected facilities. If nongrid production is also eligible for the RPS, compliance verification will be much more difficult and may require some other entity.

2.5.4 *Setting and Collecting Penalties for Noncompliance*

Penalties for noncompliance may be set at senior levels of government, for consistency with other public policies, but could then be administered by the grid operator or environmental agency.

3. EARLY ADOPTERS

The renewable portfolio standard is a new policy tool. Many jurisdictions (national, regional) are studying it, and several have RPS legislation under consideration. But as of 2003, only a few jurisdictions (and only in industrialized countries) have enacted RPS legislation. Although it is too early to assess the effectiveness of the RPS relative to alternative policies, a survey of early adopters provides an initial indication of the preferences of policymakers when trading-off implementation objectives and constraints. Because the RPS legislative situation is so dynamic, this survey is limited to general policy developments, and even some of these may soon change.

3.1 United States

The United States Congress has considered several RPS legislative proposals. In the spring of 2003, the U.S. Senate passed a bill requiring new renewables to provide an additional 1% of U.S. electricity by 2005 and 10% by 2020. The RPS would include a credit trading mechanism between producers to reduce the cost of compliance. The bill also sustains a production tax credit for renewable generators. In 2003, just over 10 U.S. states had enacted an RPS, but many additional states were considering legislation. States with an RPS enacted or in final consideration included Maine, Massachusetts, Connecticut, Pennsylvania, New Jersey, Minnesota, Wisconsin, Iowa, Texas, New Mexico, Arizona, Nevada, and California. Most RPS legislation has been associated with electricity sector reform toward greater generation competition, with RPS targeting electricity retailers. Almost all programs envision flexibility mechanisms (tradable credits), although these are mostly in the development stage. State-level electric utility regulators are usually charged with the responsibility for program administration. Urban waste and landfill gas are often included as eligible resources, and sometimes technologies such as fuel cells and cogeneration.

Variation in the RPS from one state to another reflects the unique circumstances and objectives of each state. At 30%, Maine's RPS target appears ambitious, but the RPS includes all currently operating renewables and the state already generates 50% of its electricity from eligible hydro and biomass resources. RPS advocates hail the Texas RPS as the best-designed and most ambitious policy in the

country. It envisioned the development of an additional 2000 MW of mostly wind power by 2009, which, in combination with existing renewables, would provide about 3% of total state electricity provision. Wind power has expanded rapidly since the inception of the RPS.

3.2 European Union

As part of its effort to establish a liberalized internal electricity market, the EU has been developing harmonized rules on the treatment of renewables in order to avoid distortion of trade and competition in electricity from renewable sources. In 1997, the EU issued a White Paper setting a goal for a minimum 12% share of renewable electricity supply by 2010. In 2000, the EU issued a directive proposal for a community-wide policy. To date, the directive proposal is not binding, nor does it specify the policy that should be used. Nonetheless, interest in the RPS at the EU-wide level is strong, and several countries have implemented the policy and several are close to implementation.

In Italy's RPS, operators who produce and import more than 100 GWh are obliged, from 2001, to provide at least 2% of the electricity produced and imported (net of cogeneration, exports, and self-consumption) from new, renewable generation. Renewables generators earn green certificates, which they can use or trade to nonrenewables producers in a tradable certificate market. The goal of the RPS is to triple electricity production from renewable sources by 2010.

The government of The Netherlands initially negotiated a voluntary RPS with the country's major utilities, but participants behave as if it were mandatory because meeting their RPS quota allows them to avoid an environmental tax. Producers of renewable electricity (utilities and independent power producers) receive for each kilowatt-hour provided to the grid a subsidy, which is financed by an environmental tax throughout the country on conventional energy consumption. Producers and utilities can also trade green energy certificates among themselves to meet their obligation under the RPS.

England has recently implemented an RPS with tradable certificates (called the Renewables Obligation). This replaces its system of the past decade, which involved competitive bidding by potential renewables generators for funds collected from all customers in a tariff charge called the Non-Fossil

Fuel Obligation. That system did not stimulate as much renewables production as did the feed-in tariffs for producers of Denmark, Germany, and Spain, but it did lead to declining bid prices for electricity offers to the grid from renewables producers. Sweden, Belgium, and Austria implemented RPS policies in 2003 and several other European countries are close to implementation. Denmark has been known for its generous feed-in tariffs that have fostered the dramatic growth in wind power in that country. But the subsidies are a substantial drain on public resources (100 million Euros in 1998), and the Danish parliament recently developed legislation to create an RPS, slated for implementation in the 2004–2005 period.

3.3 Australia

The Australian federal government enacted a nationwide RPS in 2001 to operate within its competitive electricity market. The policy includes a certificate trading scheme.

4. PRELIMINARY ASSESSMENT

The RPS is a policy tool to expand production of renewable-generated electricity to reflect what its role would be if long-run social costs instead of short-run financial costs were the determinant of electricity generation investment. By establishing a guaranteed market share for renewables producers, the RPS provides a strong incentive for new renewable investment, an ongoing motivation for current and new renewable producers to reduce costs, and a relatively sure means of attaining environmental targets while reducing the requirement for government financial and managerial involvement. The RPS can be applied both to electric utilities that are regulated monopolies and to competitive independent producers, but it is generally associated with jurisdictions that have reformed their electricity sector to increase competition in electricity generation. Several European countries, several U.S. states, and Australia have thus far implemented the RPS, but it is under active consideration in many other jurisdictions. Legislation and implementation are underway for an EU-wide RPS in Europe and a federal-level RPS in the United States.

It is too early to provide an empirical evaluation of the RPS relative to alternative mechanisms. However, the survey reported in here provides

information on the critical considerations in designing an RPS and a general sense of the choices being made by different jurisdictions. A few generalizations can be offered at this early stage in the development of the RPS.

The selection of the RPS target involves a three-way trade-off between environmental improvement, long-term technology strategy, and short-run cost. Jurisdictions that are especially focused on getting the greatest environmental improvement at the least cost will have only one standard for all renewables and may even include technologies that are not renewables. Jurisdictions that are interested in long-term development of advanced renewables technologies will tend to have categories of renewables (and perhaps environmentally desirable technologies that use fossil fuels), ensuring a more diverse mix of renewables and associated clean technologies. Most jurisdictions are concerned with the uncertainty of cost, so cost caps or equivalent mechanisms are common.

In terms of eligibility, most jurisdictions are starting out conservatively, limiting the scope to local renewables, excluding off-grid renewables, and trying to make sure that truly new renewable resources result from the RPS. Geographic application of policies is usually limited to individual electricity jurisdictions. However, with Australia's federal system as an example, there are efforts at the U.S. federal level and the EU level for multi-jurisdictional application. The general tendency is to apply the standard to providers rather than to consumers of electricity. The standard is almost universally applied to energy production rather than to capacity, which is consistent with environmental objectives.

Flexibility mechanisms are widely accepted as desirable. Their implementation is slowed by the administrative complexity that they can entail. Administration of the RPS in the United States and Australia is conducted through a combination of government and delegation to electricity regulators. In Europe, it is more in the domain of government at this point, but various institutional reforms are being undertaken. Thus far, all jurisdictions implementing a RPS have continued with at least some of the other policies that support renewables, whether these be research and development and demonstration projects, tax credits, or subsidized feed-in tariffs. This final point reintroduces the more fundamental concern about how the RPS compares to alternative policies, given the public objectives it was meant to serve. Critiques have argued that the

RPS is inevitably flawed because it forces the adoption of particular forms of energy, even though these are but indirect means to an end, such as environmental improvement. In so doing, the RPS may delay or prevent the development and commercialization of technology innovations that would more economically achieve the same objective. If, for example, a key motive for the RPS is to reduce greenhouse gas emissions, a policy focused directly on these emissions—such as a greenhouse gas tax or emission cap and tradable permit system—would open the door to technologies such as the production of hydrogen from coal with capture and permanent storage of all potentially harmful by-products in geological formations. Conceivably, this or other alternatives to renewables could meet all of the public objectives behind the creation of the RPS at a much lower cost, freeing resources for other social objectives.

Advocates of renewable energy have countered that nonrenewable resources such as “clean coal” cannot possibly outperform renewables in providing affordable energy while also satisfying long-term environmental, security, and social objectives. A less exclusionary argument in favor of the RPS is that renewables are likely nonetheless to be part of a future energy mix that might one day also include cleaner fossil fuel and nuclear technologies, but that renewables need special help at this stage to pass critical commercialization thresholds.

In particular, the RPS has several attractive features for driving long-run technological change in renewables. It can provide an incentive that influences the contribution of renewables to new electricity generation investment without changing the costs and economic viability of existing facilities; thus, it produces technological change at a pace consistent with the natural rate of equipment stock turnover and does not force economic obsolescence on existing equipment. For consumer prices, the higher costs of new renewables are blended with the lower costs of existing, conventional generation; price changes are therefore negligible and the likelihood of political acceptability is greater. The RPS ensures that renewables attain critical commercialization thresholds that should enable them to achieve economies of learning and economies of scale, with their costs of producing electricity falling accordingly. If the RPS includes tradable certificates and other flexibility mechanisms, it reduces the policy cost and again increases the likelihood of political acceptability. The RPS can be closely linked to environmental objectives because it provides a higher

level of confidence of the environmental outcome than do price-based policies. The RPS does not require government subsidies, and can even allow a reduced role for government in the policy's administration, because this role is delegated to electricity regulators. Finally, RPS targets can be set and adjusted to ensure that the costs of renewables expansion are not significantly greater than those of alternative means of achieving the slate of policy objectives they are intended to satisfy.

SEE ALSO THE FOLLOWING ARTICLES

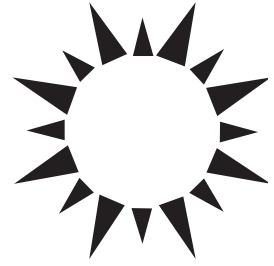
Biomass for Renewable Energy and Fuels • Computer Modeling of Renewable Power Systems • Public Reaction to Renewable Energy Sources and Systems • Renewable Energy and the City • Renewable Energy in Europe • Renewable Energy in Southern Africa •

Renewable Energy in the United States • Renewable Energy Policies and Barriers • Renewable Energy, Taxonomic Overview

Further Reading

- Berry, T., and Jaccard, M. (2001). The renewable portfolio standard: Design considerations and an implementation survey. *Energy Policy* 29, 263–277.
- Loiter, J. M., and Norberg-Bohm, V. (1999). Technology policy and renewable energy: Public roles in the development of new energy technologies. *Energy Policy* 27, 85–97.
- Menanteau, P., Finon, D., and Lamy, M.-L. (2003). Prices versus quantities: Choosing policies for promoting the development of renewable energy. *Energy Policy* 31, 799–812.
- Voogt, M., Boots, M., Schaeffer, G., and Martens, J. (2000). Renewable electricity in a liberalized market—The concept of green certificates. *Energy Environ.* 11, 65–80.
- Wiser, R., and Langniss, O. (2001). The renewable portfolio standard in Texas: An early assessment. Report No. LBNL-49107. Lawrence Berkeley National Laboratory, Berkeley, California.

Reproduction, Energy Costs of



THOMAS H. KUNZ

Boston University
Boston, Massachusetts, United States

KIMBERLY S. ORRELL

Virginia Polytechnic Institute and State University
Blacksburg, Virginia, United States

-
1. Introduction
 2. How Energy Is Expended for Reproduction
 3. Techniques Used to Study Reproductive Energetics
 4. Reproductive Energetics of Reptiles
 5. Reproductive Energetics of Birds
 6. Reproductive Energetics of Mammals
 7. Concluding Remarks

Glossary

- altricial** When offspring are hatched or born at an early stage of development.
- basal metabolic rate (BMR)** Resting metabolic rate of a homeotherm measured at thermoneutrality.
- energy budget** Partitioning of energy expenditure into the competing demands of maintenance, growth, reproduction, and storage.
- gestation** Period of embryonic development in live-bearing vertebrates.
- iteroparous** More than one reproductive event per lifetime.
- lactation** Period of milk production in mammals.
- life history trade-offs** Constraining relationships among life-history traits (e.g., current reproduction and survival, current reproduction and future reproduction, number and size of offspring).
- life history traits** Traits that define the life cycle of an organism (e.g., size at birth, growth pattern, mode of reproduction, age and size at reproductive maturity, size of offspring, fecundity, reproductive life span, longevity).
- oviparous** Egg laying; offspring develop within eggs outside a female's body.
- precocial** When offspring are hatched or born at an advanced stage of development.
- production energy** Energy allocated to growth (e.g., lean somatic tissue), storage (e.g., lipid deposition), and reproduction (e.g., gonadal recrudescence, gametogenesis, egg production).

reproductive effort Proportion of total energy devoted to reproduction within a biologically meaningful time interval.

reproductive success Number of offspring produced by an individual.

semelparous One reproductive event per lifetime.

viviparous Live-bearing; offspring develop fully or partially within the female reproductive tract.

The amount of energy that an animal can devote to reproduction is limited. The habitat in which an animal lives may contain limited food and/or water resources (e.g., deserts, high altitudes), environmental conditions may limit energy acquisition by reduced foraging activity (e.g., predation, daily or seasonal weather extremes), and there are morphological and physiological constraints on the rates at which food can be obtained and processed. Within an individual, energy must also be allocated to physiological processes and behavioral activities associated with survival and growth (e.g., foraging, thermoregulation, predator evasion). Thus, the total energy that an animal obtains must be allocated in such a way that balances maintenance, growth, reproduction, and survival.

1. INTRODUCTION

Life history traits largely determine how the energy intake of an animal is allocated to reproduction, growth, and survival. Some life history traits include the size of an animal at birth (or at hatching), the age and size at which an animal reaches reproductive maturity, how it reproduces (oviparous or viviparous), how often it can reproduce (semelparous or iteroparous), the length of time it is reproductively

active, the number and size of offspring produced, and its expected age-specific mortality and longevity. Constraints or trade-offs exist among life history traits, many of which are physiological and relate to energy management. A classic example of such a trade-off is size versus number of offspring. The allocation of energy by an adult female to produce larger offspring may improve the offspring's chances of survival, but it also can reduce the number of offspring that the female can produce. Another trade-off is between the energy allocated to current reproduction and future survival. If an animal allocates too much energy to reproduction, not enough energy will be available for maintenance or vice versa. After fat reserves are depleted, other body tissues (muscle mass) are metabolized for energy and body condition declines. Deterioration in body condition and immune function contributes to susceptibility to predators, parasites, and disease, eventually leading to death. In semelparous species (e.g., salmon), this is exactly what happens; an individual allocates all available energy to a single burst of reproduction and then dies (termed "big bang" reproduction). Iteroparous species are those that have a long enough life span that they have a high probability of surviving to reproduce more than once, and their life history traits tend to balance energy allocated to current reproduction and maintenance. Iteroparous species also illustrate another important life history trade-off: current reproduction versus future reproduction.

Longevity, age-specific mortality schedules, and reproductive life span also have a strong influence on the patterns of energy allocation to reproduction as individuals age. For example, some species invest less energy in reproduction when they are young and relatively small in body size than when they are older and larger. Smaller individuals have fewer energy reserves than do larger individuals, they have less space in their body cavity (limiting the number of eggs or offspring they can produce), and they have a narrower pelvic canal that limits egg or offspring size. Thus, smaller individuals are constrained both energetically and anatomically and, thus, produce fewer and/or smaller offspring rather than more and/or larger individuals. By allocating less energy to reproduction, smaller individuals can allocate more energy to growth—building up body condition and attaining larger body size more quickly than if they allocated less energy to growth. When individuals are older and larger, they can better afford to allocate a greater proportion of their energy to reproduction and do so when it enhances their reproductive

success. The question arises as to why young and small iteroparous animals should reproduce at all. Why not allocate energy to growth and delay reproduction until they are larger? In fact, many species do delay reproduction when they are young, even though they are physiologically mature and able to reproduce. The age-related schedule of energy allocation to reproduction in a given species is part of a suite of life history traits that coevolve to optimize the balance of energy allocated to maintenance, growth, reproduction, and survival.

Reproductive energetics is also a central aspect of theories that explain sexual selection and the evolution of mating systems. Bateman's concept of anisogamy (differential investment by each sex in gametes) and Triver's concept of parental investment are at the root of the relationship among reproductive energetics, sexual selection, and the evolution of mating systems. In general, females allocate energy to relatively few, but energetically expensive, eggs. In some species, females make a substantial energy investment in gestation, lactation, and other forms of parental care. Thus, the reproductive output of females is limited primarily by energy resources. Because variance in female reproductive success is relatively low, the influence of sexual selection on females is relatively weak. In contrast to females, males allocate minimal energy for the production of sperm; thus, male reproductive success is not usually limited by the amount of energy available for gamete production. Instead, male reproductive success is limited by the ability to compete for and successfully mate with females. As a consequence, males often invest a considerable amount of energy in activities and traits associated with mating success (e.g., advertisement displays, territorial defense, agonistic encounters, courtship, large body size, structures used for courtship or intermale combat). In addition, males that are emancipated from parental care typically invest time and energy in defending or pursuing multiple mates, contributing to a polygynous mating system, and engaging in intense intermale competition for mates. For other species, both males and females (or, in rare cases, males alone) are required to invest time and energy in caring for offspring. Caregiving males rarely have the time and energy to pursue multiple mates; thus, a monogamous mating system usually results. Although monogamous males also expend energy to compete for mates, their reproductive success does not vary as much as that of polygynous males; thus, intermale competition and the influence of sexual selection on males are not as intense as in polygynous systems.

2. HOW ENERGY IS EXPENDED FOR REPRODUCTION

Because of wide variation in life history traits, reproductive biology, and mating strategies, the ways in which energy is expended for reproduction differ greatly among species and between the sexes. Because most vertebrate species breed seasonally, they do not typically produce gametes year-round, and many species maintain nonreproductive gonadal tissues in a reduced state. Thus, the initial investment of energy into reproduction actually begins prior to mating when energy is expended for gonadal recrudescence. Once hormones are produced, they stimulate the development or enlargement of the testes, ovaries, and oviducts or uterus. Once the reproductive organs are ready, energy is expended to produce gametes. For males, this includes spermatogenesis and production of seminal fluids. For females, egg production starts with oogenesis but goes beyond the simple production of sex cells. Oviparous reptiles and birds allocate a considerable amount of energy to egg production; eggs are provisioned with a large mass of energy-rich lipids or yolk, egg white (water and albumen), and a calcified eggshell. In comparison, viviparous (live-bearing) mammals produce eggs that contain little stored energy but allocate a large amount of energy to carry and nourish developing embryos—and for lactation to nourish offspring for all or part of the period of postnatal growth.

In addition to physiological investment in reproduction, a large amount of energy is expended for reproductive activities. Examples of reproductive activities include finding and defending a suitable territory or nest site, fighting or repelling competitors, collecting nest materials and building a nest or excavating a burrow, and finding, choosing, and courting a mate. The act of copulation is a reproductive activity, as is the laying of eggs, giving birth to live offspring, and guarding, burying, incubating, or carrying eggs. Males of some species actively guard their mates prior to and/or after copulation to ensure paternity. For crocodilians as well as many bird and mammal species, reproductive activities in the form of parental care may continue long after offspring are hatched or born. These activities include feeding the offspring while they are unable to forage for themselves, guarding them from predators, and tutoring them in social and survival skills. In some cases, it may be difficult to draw the line as to whether a particular activity should be classified as reproductive. Nonbreeding types of

activities that occur during the breeding season may occur more frequently to support reproduction. For example, some animals might need to forage and feed more often prior to or during the breeding season to support reproduction, and males might forage to provide nuptial gifts to impress prospective mates or to feed nest-bound mates that incubate eggs or guard offspring. Many animals (mostly birds and mammals) exert some degree of thermoregulation to survive, but reproductive females might need to maintain higher than normal body temperatures and basal metabolic rates (BMRs) to promote the optimal development of offspring or to increase the rate of digestion, assimilation, and egg or milk production.

3. TECHNIQUES USED TO STUDY REPRODUCTIVE ENERGETICS

Because energy is expended for reproduction in so many different ways, quantifying this energy is often complicated and difficult. As a result, many studies examine only part of the energy expended for reproduction. One of the easiest components of reproduction to determine is the energy allocated to the storage of fat and the production of eggs. Lipid extraction is used to separate body fat from lean tissue and to quantify the energy contained in body fat. Bomb calorimetry is used to determine the energy content of eggs and lean body tissues (Table I). Changes in body mass and body composition observed in a cohort of individuals over time, and the energy content of fat and lean tissue, can be used to calculate the amount of energy allocated to storage and/or growth (Fig. 1).

Another technique used to examine reproductive energetics is laboratory respirometry (Table I). Respirometry is used to measure metabolic rates (in milliliters O₂ consumed per hour or CO₂ produced per hour) of inactive animals held at species-typical body temperature(s). BMRs (in mammals or birds) and standard metabolic rates (SMRs) (in reptiles) of fasted sleeping animals are the lowest rates of energy metabolism and represent the amount of energy expended for basic body functions. Resting or maintenance metabolic rates (RMRs) are determined in inactive or sleeping animals digesting food and so include the cost of digesting and assimilating food (specific dynamic action). Some studies have determined metabolic rates of animals engaged in various activities. Laboratory metabolic rates have frequently been used to extrapolate the amount of total energy expenditure, with extrapolated values then being used

TABLE 1

Analytical Techniques Used to Study Reproductive Energetics of Terrestrial Vertebrates

Activity profiles: Time budgets detailing the amount of time spent in various activities. They may be combined with respirometry to calculate the amount of energy allocated to activity and maintenance.

Lab respirometry: Measures the metabolism (in milliliters O₂ consumed and/or CO₂ produced per hour, converted to joules of energy) of inactive animals under laboratory conditions.

Doubly labeled water: Determines the daily energy expended or consumed (in milliliters CO₂ per day, converted to joules of energy) by free-ranging animals. Isotopes of oxygen (O-18) and hydrogen (deuterium or tritium) are injected into an animal, blood samples are taken over two or more intervals of time, and then the turnover rate in isotopes is used to calculate the amount of energy consumed during that time by oxidative metabolism.

Lipid extraction: Used to separate and quantify the energy contained in whole body lipids. An animal carcass is dried to a constant mass and macerated, and solvents are used to extract the lipids for further analysis.

Bomb calorimetry: Determines the energy content of organic substances (e.g., body tissues, eggs). Dried material is incinerated in high-pressure oxygen in a stainless steel container at a constant volume. The heat produced by the reaction is absorbed by water surrounding the container, and the change in water temperature indicates the amount of energy released.

Energy budgets: Where data from several different techniques are combined to partition total metabolizable energy into that allocated to competing categories of resting maintenance, activity maintenance, and production.

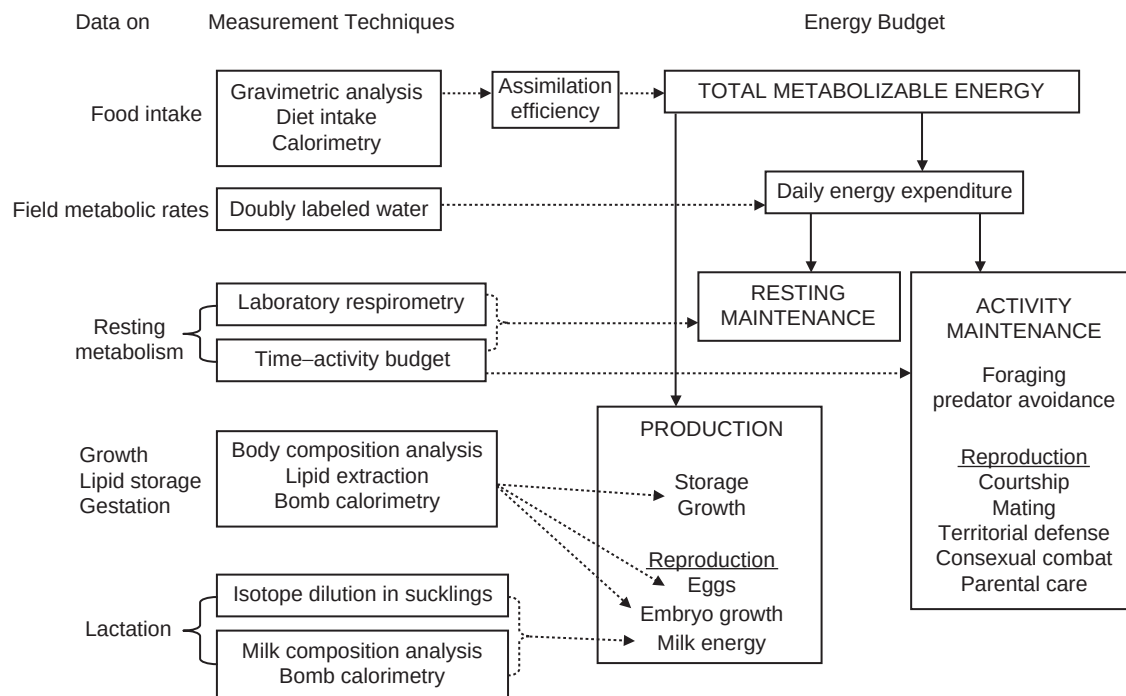


FIGURE 1 Schematic diagram of a generalized energy budget used to examine reproductive energetics, including the type of data used, measurement techniques and definitions of energy budget components, and methods used to construct an energy budget. Modified from Kunz and Nagy (1988)

to construct energy budgets (Fig. 1). Similarly, RMRs combined with time-activity budgets can be used to estimate the energy allocated to maintenance or resting metabolism for a fixed interval of time.

The doubly-labeled water (DLW) method (Table I) is used to determine the total daily energy expenditure (DEE) (in kilojoules per day) of free-ranging

animals under field conditions. The DLW method is the only energy measurement technique that includes the energy allocated to activity under natural conditions. Thus, the DLW technique can be used in combination with other techniques (Fig. 1) to examine the energy allocated to reproductive activities.

Energy budgets combine several techniques to partition total energy expenditure into compartments of maintenance, growth, reproduction, storage, and export (as milk). The construction of an energy budget requires several different techniques (Table I): (1) the DLW method to determine daily energy expenditure, (2) laboratory respirometry and time-activity budgets to calculate energy expended for maintenance, (3) bomb calorimetry to measure the energy content of eggs (for birds and oviparous reptiles), (4) quantification of energy accumulated prior to breeding (fat), and (5) the amount of energy exported as milk by mammals. If one can quantify the total energy expended by an animal, the energy expended for maintenance metabolism, the energy allocated to egg production in reptiles and birds and to embryonic growth and milk production in mammals, and energy allocated to growth, the energy that is allocated to activity can be estimated by subtraction. The energy expended for reproductive activities can then be estimated based on time-activity profiles and the amount of energy stored in fat or exported in milk.

A difficult part of examining reproductive energetics is determining the appropriate temporal schedule of reproduction. Although a few species are semelparous and reproduce only once before they die (e.g., salmon, spiders, many insects), this is not a typical reproductive pattern in vertebrate animals. Most species are iteroparous (reproducing more than once) and may reproduce once per year for several years in succession or even multiple times per year, whereas others may skip years. A reproductive period or season may be relatively short for some species (a few weeks or months), whereas other species may spend several months or years providing parental care to ensure the survival of their offspring (e.g., alligators, primates, elephants). In some iteroparous species, the amount of energy allocated to reproduction varies greatly with age. Thus, for some species, it may be necessary to investigate the reproductive energetics of a cohort of individuals over their entire life span.

4. REPRODUCTIVE ENERGETICS OF REPTILES

4.1 Territorial and Courtship Activities

There are many different types of mating strategies exhibited by reptiles, but most fall within two basic patterns of behavior: territorial (most lizards and

crocodilians) and nonterritorial (most snakes and turtles). In territorial species, one or both sexes (usually males) actively defend a fixed area to monopolize the resources (food, shelter, or nest sites) and/or the mates (usually females) residing within it and to prevent intrusion or takeover by consensual competitors. Territoriality should evolve when the benefits (e.g., access to resources such as space, food, and females) exceed the costs incurred (e.g., energy expenditure, risk of injury). A large amount of energy is expended for territorial activities, including visual and/or chemical signals used to advertise the territory holder's presence and territorial boundaries, aggressive interactions and/or combat with consensuals, and territorial patrol to detect and repel intruders.

Nonterritorial reptiles may stay within a well-defined or loosely defined home range that is not defended, or they may roam widely in search of mates. Although nonterritorial reptiles often use chemical signals to find mates, they do not typically invest energy in the conspicuous and frequent visual signaling and patrol activities characteristic of territorial males. However, it is typical of both territorial and nonterritorial reptiles to invest energy in aggressive conflicts and combat with consensuals, especially if males are competing for a territory or a female.

During agonistic encounters, male lizards of many species typically perform head-bob or push-up displays accompanied by size-enhancing displays and body postures, mouth gapes that convey the threat to bite, and ritualized circling and jaw sparring with opponents. Males in some snake species also exhibit ritualized combat. Male rattlesnakes wrap around each other and raise their upper bodies off the ground, and then each attempts to wrestle and pin its opponent's head to the ground. Even some male turtles (e.g., snapping turtles, wood turtles, gopher tortoise) inhabiting the same area may fight each other to establish their position in a dominance hierarchy to gain the privilege of mating with females.

Most reptile species invest energy in some type of courtship behavior. Because females are typically choosier than males about their mating partners, it is usually the males of the species that do most of the courting in an attempt to persuade females to mate with them. Courtship in reptiles typically consists of species-specific visual displays (e.g., lizard head-bob displays, display of colored skin patches), acceptance or rejection behaviors (usually given by females), and/or tactile behaviors (e.g., licking, stroking, biting, chin rubbing) exchanged between the sexes. Reptiles do not exhibit the elaborate courtship rituals often observed in birds.

The amount of energy that reptiles allocate to copulatory activity varies greatly with the type of mating system. Some reptile species copulate only once or a few times during a breeding season, whereas others copulate repeatedly, especially polygynous males. For example, males of the green anole (*Anolis carolinensis*) expend energy to court and mate with an average of three females residing within their territory, mating repeatedly with each female roughly once each week for the duration of a 4-month breeding season. Because the purpose of territorial behavior is to exclude male competitors, territorial males do not usually guard their females directly after mating. Nonterritorial species may be polygynous, promiscuous (both sexes mate with multiple individuals), or even monogamous. In some nonterritorial species, males expend energy to guard their mates following copulation to prevent them from mating with other males. Guarding their mates may also preclude males from mating with other females (hence, they are monogamous), and a few species form long-term pair bonds that span multiple reproductive seasons (e.g., snow skinks [*Niveoscincus microlepidotus*], sleepy lizards [*Tiliqua rugosa*]). The duration of copulations, and thus the amount of energy allocated to copulation, also varies widely,

even among closely related species. Some *Anolis* lizards have copulations lasting for several seconds, whereas males of *A. carolinensis* average approximately 25 to 30 min for a single copulation. In many vertebrate taxa, the duration and frequency of copulations are related to the intensity of sperm competition in a species, although studies addressing this issue in reptiles are currently lacking.

Four studies on lizards have quantified the energy expended for activities during the breeding season, and all of the species examined are territorial and polygynous (Table II). For a single reproductive season, the proportion of energy expended for activities during the breeding season ranged from 23 to 57% of total energy expended during that time. For males of *A. carolinensis*, nearly all of their activity is devoted to reproduction (e.g., only 1–2% of their day is spent foraging) and nearly half of their total energy expenditure is allocated to reproductive activity (Table II).

4.2 Gamete Production

Reptiles are similar to most vertebrate species in that males typically allocate little energy to producing sperm. The amount of energy that gonadal recrudescence and sperm production requires has not been

TABLE II

Comparison of the Energy Expended for a Single Reproductive Season in Several Lizard Species (All Figures in Kilojoules per Season Unless Otherwise Noted)

Species	Mating system ^a	Sex	Body mass (g)	Daily energy				
				expenditure (kJ)	Eggs (kJ)	Reproduction (kJ) ^b	TME (kJ) ^c	TME(kJ/g ^{0.8}) ^d
<i>Anolis carolinensis</i>	T, P	M	5.7	154	—	74	154	38.3
		F	3.3	105	59	109	164	63.1
<i>Callisaurus draconoides</i>	T, P	M	10.5	168	—	Not examined	168	25.6
		F	8.0	132	53	Not examined	185	35.1
<i>Cnemidophorus tigris</i>	NT, PR	M	18.8	507	—	Not examined	507	48.5
		F	15.8	361	52	Not examined	413	45.4
<i>Iguana iguana</i>	T, P	M	713	3320	—	1892	3320	17.3
		F	1004	4331	3847	6317	8178	32.5
<i>Sceloporus jarrovi</i>	T, P	M	23.9	174	—	40	174	13.7
		F	17.4	87	43	36	130	13.2
<i>Sceloporus occidentalis</i>	T, P	M	11.9	186	—	105	186	27.9
		F	11.9	200	32	151	232	32.3
<i>Uta stansburiana</i>	T, P	M	3.7	64	—	22	64	22.5
		F	2.4	32	18	50	45	24.8

Source. Modified from Orrell *et al.* (2004).

^aKey to mating system: T, territorial; NT, nonterritorial; P, polygynous; PR, promiscuous.

^bReproduction: energy allocated to breeding season activity by males, and to breeding season activity + eggs for females.

^cTME: Total metabolizable energy, calculated as daily energy expenditure during the breeding season + energy allocated to eggs.

^dTME/body mass^{0.8}: TME adjusted for interspecific differences in body size, to facilitate interspecies comparisons.

quantified, but it is likely to be a very small amount relative to the energy expended for other aspects of reproduction. In contrast, nearly all female reptiles allocate a considerable amount of energy to producing large eggs that contain energy-rich yolk. The size and number of eggs produced by different species varies considerably. However, the amount of energy contained in egg materials is very similar among reptile species, averaging approximately 6 calories per milligram of dry egg mass. The amount of energy allocated to eggs by lizards ranges from 13 to 48% of total energy expended during the breeding season (Table II) and from 7 to 13% of total annual energy expenditure. Similarly, the painted turtle (*Chrysemys picta*) allocates approximately 14% of its annual energy to eggs.

4.3 Nest-Related Activities, Incubation, and Parental Care

Because most reptiles produce precocial and independent offspring that develop and hatch from shelled eggs (ovipary), reptiles rarely allocate energy to parental care activities. Instead, reptile species provide a form of nutritional parental care by provisioning their eggs with “extra” yolk that nourishes the hatchling for days or weeks after emergence from the egg. Female reptiles may also allocate energy to digging holes or nest cavities in which they lay eggs and then cover them with substrate (e.g., most turtles, many lizards). In crocodilians and some snake species, females construct a nest mound made from soil or vegetation in which to lay eggs, where decaying organic matter provides heat for optimal incubation temperatures. In species with relatively small eggs and/or clutches (e.g., *Anolis* lizards lay single-egg clutches), burying eggs may require a relatively minor amount of energy relative to other reproductive activities. In species with large eggs, large clutch sizes, and/or long travel distances to find nest sites (e.g., sea turtles travel many miles to lay eggs on natal beaches), the energy expended for nest-related activities may constitute a substantial energy investment.

In addition to nest construction, some reptile species guard their nest of eggs for all or part of their development (e.g., king cobras, crocodilians, pythons, some lizards). Female alligators construct and guard their nest of eggs and then protect their young for up to 2 years. Female Burmese pythons (*Python molurus*) actually incubate their eggs by coiling around them and rhythmically contracting their body muscles, generating metabolic heat to

maintain the optimal incubation temperature of 30°C. The increase in energy expenditure associated with incubation for female pythons is 2 to 20 times higher than that for females that are not incubating, with an inverse relationship between ambient temperature and the rate of energy expended; that is, it takes more energy to keep the eggs warm at cooler ambient temperature.

4.4 Egg Retention and Gestation

Most reptiles (chelonians, crocodilians, tuataras, and most lizards and snakes) are oviparous. They lay eggs with relatively undeveloped embryos and a large yolk mass containing enough energy to support embryonic development. For the most part, oviparity excludes the energetic costs of gestation. However, as many as one-fifth of squamate species (snakes, lizards, and amphisbaenids) exhibit various degrees of viviparity that require some amount of energy expended for gestation. At one end of the ovipary/vivipary continuum, there are many oviparous squamates that retain developing eggs in the oviducts for some length of time. Further along this continuum, ovoviviparous females provide eggs with yolk for embryo development, but eggs are enclosed by a noncalcified shell or membrane and remain in the oviduct until completely developed (e.g., *Boa constrictor*). Various intermediate states include some species providing both a yolk mass and some type of placental nutrition of embryos (e.g., garter snake [*Thamnophis sirtalis*]) or a highly vascularized oviduct (e.g., common lizard [*Lacerta vivipara*]). At the other end of the continuum, a few species exhibit extremely reduced yolk mass, and offspring are attached and nourished by a placenta formed from uterine tissue and extra-embryonic membranes (eu vivipary) (e.g., skinks [*Mabuya heathi* and *Chalcides chalcides*]).

In addition to nutrition that may be provided directly to offspring, there are a number of energetic costs for egg retention and embryo gestation that do not occur with strictly oviparous species. Egg-retaining/Gestating females must expend energy to behaviorally thermoregulate and maintain optimal body temperatures for developing embryos. In addition, if egg-retaining/gestating females maintain higher than normal body temperatures (and studies show that many do), this in turn increases their own temperature-dependent maintenance metabolism. Energy is also expended to carry around the bulk and mass of the eggs and/or developing embryos. Other types of costs for egg retention and embryo gestation include decreased mobility and greater

susceptibility to predators, smaller offspring and/or clutch sizes, and lower rate of offspring production. Given all of the potential costs, there must be tremendous selective pressure on some species to evolve toward some degree of viviparity. The potential benefits of egg retention or gestation include optimal temperatures for embryonic development, delayed oviposition (or birth of live offspring) when environmental conditions (e.g., moisture, temperature) are unfavorable to embryonic development within eggs, and protection of eggs from predators, bacteria, and fungi. There is a strong trend for egg retention and viviparity to occur in squamates at high elevations, where extremes in temperature, humidity, and low atmospheric oxygen concentration inhibit or preclude embryonic development if eggs were subject to these conditions. Clearly, egg retention and various forms of gestation have both costs and selective benefits in terms of energy allocation and mortality risk versus increased offspring survival and reproductive success, making the evolution of egg retention and viviparity an example of another life history trade-off.

5. REPRODUCTIVE ENERGETICS OF BIRDS

5.1 Territoriality, Courtship, and Mating

The vast majority of bird species (>90%) have a monogamous mating system in which individuals (typically males) compete with each other for access to mates and resources critical to mating. When mates or resources are economically defendable, individuals monopolize mates and/or resources through territorial behavior. The area defended may be large and contain mates or resources important to reproduction (e.g., food, mating, nest sites), or it may be small and encompass only the space immediately surrounding the nest to protect the offspring, nest, nest site, and/or nesting materials (e.g., colonial nesting species). Song is one of the most conspicuous and energetically demanding aspects of territorial behavior of birds. Birds (mostly males) use species-specific songs to advertise their presence in a territory, to repel potential intruders, to find and attract potential mates, and to stimulate female reproductive physiology and behavior. Studies have shown that normal males rendered mute experimentally are far less effective at repelling intruders and/or maintaining their territory and mates than are vocal males. Laboratory respirometry

was used to determine the energy expended for singing in captive male Carolina wrens (*Thryothorus ludovicianus*). Their metabolic rates while singing were 2.7 to 8.7 times higher than their RMRs. Another example of energetically demanding territorial behavior is the rapid and loud drumming of woodpeckers (males rapidly tap their beaks on hollow trees or other resonant structures) and grouse (males beat their wings in the air or on their breasts while standing on logs, boulders, or tree roots).

In addition to advertising a male's presence in a territory, the frequency and/or intensity of bird song or drumming may indicate male body condition and/or energy reserves and may be a signal used by males to assess competitive ability or used by females to make mate choice decisions. In addition to singing and drumming, some birds expend energy for territorial patrol of their territory to detect and repel conspecific and/or heterospecific intruders, using visual threat displays and/or chasing to protect resources and/or offspring. When threat displays and chasing fail to oust intruders, escalated aggression can involve direct combat and the risk of physical injury. It has been estimated that some males expend up to a third of their daily energy expenditure, or up to 2.4 times the cost of BMR (in house martins [*Delichon urbica*]), for territorial activity. Perhaps because territorial advertisement is energetically costly and attracts the attention of predators, females of many bird species do not sing or call for territorial reasons or they do so much less frequently than do conspecific males. Although females do not advertise their territory, they too may chase off conspecific females to protect their offspring or resources important to reproduction.

As a result of competition among males for mates, sexual selection results in male birds expending energy for conspicuous traits and behaviors used to influence female mate choice. Males of many bird species exhibit colorful and/or modified feathers and/or structures (e.g., "red-wing" blackbirds, yellow- or orange-crested cockatoos, multicolored macaws, greatly elongated and colorful tail feathers of peacocks, wattles and combs of roosters). The conspicuous ornaments of male birds may require a minor amount of energy to produce relative to total energy expended for reproduction, but these traits may also be costly to males because they attract the attention of predators or reduce a male's ability to evade predators (e.g., long tail feathers). In addition, the quality and/or size of these traits may reflect the energy stores, body conditions, and general health of individuals and could be an important signal to male competitors and

potential mates. For instance, bright and intensely colored traits could indicate that males are free of parasites and/or disease that could otherwise compromise their ability to defend their territory or that could be transmitted to mates or offspring. Quality of male ornaments may also indicate the genetic quality of individuals, that is, parasite and/or disease resistance that will be passed to offspring or inherited traits preferred by females that confer reproductive success to those males and their future offspring.

Males expend a considerable amount of energy to show off their conspicuous ornaments, using complex and ritualized behaviors or activities to court females. Courtship displays may consist of aerial dances or ritualized sequences of behaviors performed on the ground, in trees, or on the water (e.g., head bows; fluffed feathers; erected crests; fanned, waved, or flapped wings; choreographed hops or steps), and all of these displays may be combined with vocalizations. Male bowerbirds build elaborate bowers out of vegetation, often decorating them with feathers, paper, string, or other colored objects, and then use the bowers as performance stages or backgrounds for their ritualized courtship dances. In some bird species, several males congregate in "leks" and simultaneously display and exhibit their ornaments to observing females, who evaluate them and their performances before choosing their mates. Male jays, cardinals, terns, and other species also expend energy to forage for food items presented to prospective or current mates as nuptial gifts. Courtship displays and activities often require a large investment of energy to perform, especially if performed repeatedly. The daily energy expended by male sage grouse (*Centrocercus urophasianus*) performing strut displays toward females may be twice that of nondisplaying males and positively correlated with the number of struts performed per day. In many species, the frequency or intensity of courtship behaviors is correlated with the displayer's condition (i.e., sufficient energy intake to support the activity and maintain body condition) and/or with female mating preferences. Researchers have found that female sage grouse that choose among displaying males on a lek preferred those with the highest strut rates, and a positive correlation between male foraging distance and strut rate suggests that display vigor indicates male endurance and foraging skill.

Female birds rarely exhibit the bright colors, elaborate structures, or ritualized courtship behaviors and activities observed in conspecific males, possibly because these traits are selected against in females and are costly in terms of energy expenditure and

susceptibility to predators. However, in some cases, females participate in male courtship rituals, especially when they are receptive to the courting males.

Although a single copulation is generally very brief in birds, the energy expended for copulatory behavior varies considerably. Some species copulate only a few times during the breeding season, whereas other species copulate hundreds of times. The frequency of copulations and the energy expended in this activity are related to sperm competition and paternity. Although the vast majority of bird species are socially monogamous, extra-pair copulations are common in most species. In general, the less assured a male is of paternity, the more frequently he will copulate with a female. For species in which a male guards the female before and after copulation and until her eggs are laid, the male has a reasonable assurance that the female has not mated with other males and that only one or a few copulations are required for him to fertilize her eggs. For a male that leaves his mate for extended periods of time, either to forage for food or to copulate with other females (a polygynous male or a monogamous male seeking his own extra-pair copulations), it is possible (or in some cases probable) that his female has copulated with one or more other males. Because males in many bird species invest a large amount of energy in parental care to raise offspring, it is crucial to a male's reproductive success to ensure that his sperm fertilizes his mate's eggs. Frequent copulations with a female may increase a male's paternity by inseminating a greater quantity of sperm, and/or by providing fresh sperm, to compete with sperm inseminated by other males. Studies using DNA markers have revealed that paternity in birds is influenced by copulation frequency, sperm number, fertilizing capacity of sperm, and the timing of insemination relative to other inseminations and to oviposition. Frequent copulations require greater sperm production, and many studies on birds have documented that large testes produce greater numbers of sperm than do smaller testes. Therefore, it is not surprising that copulation frequency is positively correlated with testis size in many vertebrate taxa, including birds. Because large testes may evolve in response to sperm competition, relative testis size provides a measure of the intensity of sperm competition in various species.

5.2 Gamete Production

Because most birds are seasonal breeders that produce gametes only during the reproductive season, energy

is allocated to gonadal recrudescence prior to the breeding season for both sexes. However, the increase in testes size equates to roughly 2% of the cost of male BMR, and the increase in ovary and oviduct mass constitute 2 to 9% of the cost of female BMR. Thus, females expend more energy than males for gonadal recrudescence, but for both sexes it is a relatively trivial amount compared with the amount of energy expended for other aspects of reproduction.

Although the amount of energy allocated to sperm production varies depending on copulation frequency, males still allocate a very small amount of energy to sperm and ejaculate production. For example, the energy that male chickens require to produce semen accounts for less than 1% of the amount expended for BMR. In contrast, females allocate a tremendous amount of energy to egg production. The total amount of energy allocated to egg production in a reproductive season depends on the number of eggs produced, egg size, and the amount of energy contained in the eggs, and these factors are influenced by female body size, energy availability, and the particular life history traits of the species. Although the total amount of energy allocated to egg production varies among individuals and species, the amount of energy contained in the eggs themselves generally is proportional to the developmental stage of the offspring at hatching (Fig. 2). The eggs of birds that produce altricial offspring tend to contain a relatively small proportion of yolk (24% of wet egg mass) and less energy (4.7 kJ/g of wet mass) in comparison with most

precocial species (40% of wet egg mass). The eggs of extremely precocial species (e.g., Australian megapodes such as the brush turkey [*Alectura lathamii*]) contain the greatest proportion of yolk (up to 65% of wet mass) and the greatest amount of energy (12.3 kJ/g of wet mass). Altricial offspring also tend to have the least amount of energy remaining in yolk sacs after hatching (used for development and growth [average 8% of hatchling body mass]), whereas precocial species tend to have a greater amount of energy left in yolk sacs after hatching (average 15% of hatchling body mass). The total energy allocated to egg production in various bird groups has been estimated at 40 to 170% of the energy expended for BMR.

5.3 Nest Building and Incubation

Because all birds are oviparous, nearly all species expend energy to collect nest materials, construct or excavate nests, and incubate eggs. Exceptions to this generalization include several species of parasitic birds (e.g., cowbirds) that lay their eggs in the nests of other species (or, more rarely, in the nests of conspecifics), leaving eggs to be incubated and cared for by the nest owners. Although birds invest a substantial amount of time and effort in constructing nests, the energetic cost of this effort has not been examined.

Egg incubation in birds consists of maintaining a clutch of eggs at the optimal temperature for embryonic development until they are hatched. Most birds use their own body heat to maintain egg temperatures (except megapodes, which bury and tend eggs in mounds of organic matter and depend on external sources of heat such as solar radiation, plant decomposition, and geothermal heat). The transfer of body heat to eggs and the monitoring of egg temperature are facilitated by a highly vascularized and bare patch of skin called a brood patch. In many monogamous species, each sex takes its turn incubating the eggs while the other mate forages for food. In some species, the female continually incubates the eggs while her mate provides her with food. Egg incubation may also be performed by only one parent (usually the females, although a polyandrous male may also incubate), but that parent must either forgo feeding during incubation or occasionally leave the eggs to find food. When environmental temperatures are near the thermoneutral zone of a species, the eggs may require little incubation time or metabolic heat to maintain their optimal temperature. When environmental

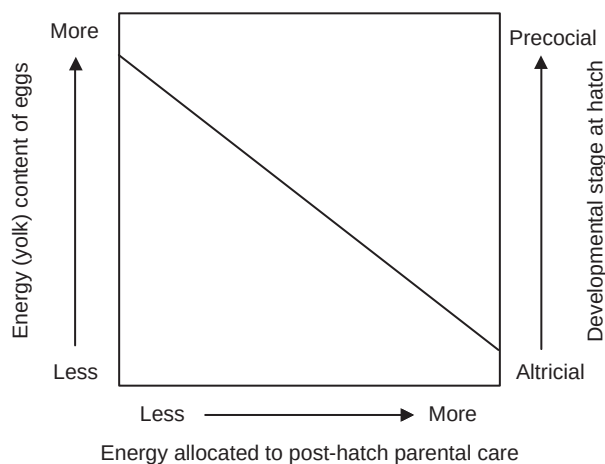


FIGURE 2 General trends in the relationships between the energy content of eggs, the energy allocated to post-hatch parental care, and the developmental stage of avian offspring at hatch.

temperatures drop below the incubating bird's thermoneutral zone, birds may have to raise their metabolism to provide enough heat to maintain egg temperature, thereby expending energy specifically for incubation. Studies indicate that incubating birds have RMRs that are 19 to 50% higher than those of nonincubating birds at the same environmental temperatures.

5.4 Parental Care

Birds invest a considerable amount of time and energy in caring for their offspring and do so in a number of different ways. Hatchlings must typically be fed, so one or both parents must find or capture enough food to support the maintenance and growth of their offspring and to support their own maintenance and reproductive activities. Parents may also process food prior to feeding it to offspring; many birds partially digest and regurgitate food, and pigeons and doves of both sexes produce crop milk secretions to nourish hatchlings. Parental care also includes brooding offspring, that is, protecting them from extremes in temperature or other types of inclement weather. Brooding is essential to offspring survival because most chicks are poikilothermic (body temperature fluctuates with environmental temperatures) for days or weeks after they hatch because they do not have the muscle mass to produce sufficient metabolic heat for homeothermy (constant body temperature) and/or because they lack insulating down or feathers to retain body heat that is produced. Thus, species that reproduce in cold regions (e.g., penguins) invest more time and energy in brooding their young than do those that inhabit warmer tropical regions. Birds often protect offspring from predators, for example, by using alarm calls to warn offspring about the presence of predators, attacking and mobbing predators to chase them away, or luring predators away by faking injuries (e.g., the broken wing act used by killdeer [*Charadrius vociferus*]). Some bird species also teach their offspring survival skills such as leading them to food sites and demonstrating how to find and capture prey and how to evade predators.

The amount of time and energy that birds expend for parental care depends on many factors, including (but not limited to) the developmental stage of hatchlings, the number of offspring in a clutch, and the abundance and distribution of food resources. To some extent, the amount of parental care provided is inversely proportional to the hatchling developmental stage; the more developed the offspring are

at hatching, the less time and energy are required for parental care. Thus, the amount of time and energy required for parental care tends to be counterbalanced by the energy invested in eggs (Fig. 2). The amount of time and energy required to raise offspring is a strong limiting factor for the number of offspring that can be raised successfully. Experiments have found that artificially increasing brood size in house martins (*D. urbica*) decreases both the size and number of offspring fledged and proportionally increases the daily energy expenditure of savannah sparrows (*Passerculus sswichensis*). An increase in energy expenditure in savannah sparrows was correlated with both brood size and the number of feeding visits to the nest. Similar experiments that increased brood size in Eurasian kestrels (*Falco tinnunculus*) found that males (which provide most of the food to offspring) were able to increase the number of voles caught to support the entire brood. However, when prey abundance was artificially reduced, male kestrels compensated by expending more energy to bring back food. When prey abundance was decreased and brood size was increased at the same time, kestrel offspring grew more slowly and had higher mortality rates, and the parents lost body mass and had increased mortality. The amount of energy that birds expend while raising broods has been determined in several species to be about three times their basal rate of metabolism.

In some species of birds, the energetic expense associated with parental care may be alleviated in part by conspecific "helpers" that provide alloparental care or assistance in raising and/or protecting the offspring of other birds. Helpers are usually related to the breeding birds. Most often, young adults from previous broods help to raise their younger siblings. Helpers usually delay their own reproduction due to limited breeding resources (e.g., territory space, nest sites, food). The breeding birds benefit from the assistance of helpers by successfully producing a greater number of offspring than they could produce on their own. In one study, the mortality of offspring in the pied kingfisher (*Ceryle rudis*) was 61% when parents were unassisted, decreased to 22% when there was one helper, and decreased to nearly 0% when there were two helpers. Helpers also benefit by gaining skill and experience in providing parental care, making them more successful at raising their own offspring later. In addition, by helping to raise related individuals, they increase their genetic contribution to the next generation.

6. REPRODUCTIVE ENERGETICS OF MAMMALS

6.1 Territoriality, Courtship, and Mating

Most mammal species (~95%) have a polygynous mating system in which males typically compete with each other for access to mates and resources. Males of polygynous species defend territories (or female groups) so as to acquire multiple mates, and the energy investment in this activity can be substantial. Males may spend a considerable amount of time patrolling boundaries of their space, vocalizing, scent marking, and making visual displays. In some instances, the displays require a relatively small investment in energy because once a territory is established, neighbors are conditioned and need only occasional reminders to keep out. Still, economic defendability is the most important determinant of when a territory or group of females should be defended. For example, male arctic ground squirrels (*Spermophilus parryii*) defend territories against other reproductively active males, and paternity analysis has revealed that the first male to mate with an estrous female sires her offspring. Male short-nosed fruit bats (*Cynopterus sphinx*) construct tents in which they roost by modifying leaves and other plant parts and then defend these roosts against intrusions of conspecific males. Different numbers of females join these males, but it is unknown whether females choose the males based on quality of the resource (tent) or on some quality of the males. Still, paternity analysis has revealed that females sometimes move among adjacent tents; thus, the resident males might not always sire the pups produced by the female occupants of their tents. If males defend females instead of some habitat, this is referred to as female defense polygyny. However, as in other mammal species, it is not always clear whether a male defends a territory or defends the female that occupies it.

Three general types of polygynous mating systems occur in mammals: resource defense polygyny, female defense polygyny, and leks. Resource defense polygyny occurs when males defend areas that contain the feeding, nesting, or roosting sites that are critical for reproduction. Female choice is generally influenced by the quality of the male and the area that he defends. Female defense polygyny commonly occurs in situations where females are gregarious for reasons unrelated to reproduction, for example, when females form herds for protection from predators (e.g., musk oxen [*Ovibos moscha-*

tus]) or when they assemble near food, nesting, or roosting sites (e.g., greater spear-nosed bat [*Phyllostomus hastatus*]). In the latter situation, females become gregarious because roost sites in caves are limited and males then monopolize the females for breeding. Polygynous mating systems usually lead to intense competition among males, and this can result in marked sexual dimorphism and large variations in reproductive success. In the greater spear-nosed bat, resident males sire from 60 to 90% of the pups born to the females in a harem.

The sac-winged bat, *Saccopteryx bilineata*, is characterized by a female defense mating system in which males defend females in a harem territory. Courtship and harem defense in this species includes energetically costly flight maneuvers and hovering displays. A doubly labeled water experiment revealed that the mass-specific field metabolic rate of harem males increased with harem size. Males typically scent mark the boundaries of their territories and defend these areas against intrusions by other males. Adult males attempt to attract females into their harem territories using acoustic, visual, and olfactory displays. In daytime roosts, which are often located in well-illuminated buttress cavities of large trees around buildings in the Neotropics, females are aggressive toward one another as well as toward males. One of the most distinguishing displays of this species is a hovering flight during which odor is fanned toward roosting females. Males perform hovering displays throughout the year, although copulations occur only during 2 months of the year. Males that roost on the periphery of harems father some of the offspring, but harem males on average are the most successful. Thus, although harem males are not the exclusive fathers of the harem offspring, male reproductive success is on average correlated with harem size.

Intense competition among males can also be observed in a lek mating system, where males congregate and display communally to attract females. The only resource that a male defends on a lek is the space where courtship and mating occurs. Males arrive early in the breeding season and, through highly ritualized combat behaviors, compete for and establish their territories. Males of lekking mammals have evolved elaborate ornaments such as horns, antlers, and vocalizations that they use in intermale competition to attract females and to influence female choice. Females move through the arena (lek) in which males are displaying and mate with one or more males based on the quality of their leks (and resources within it) or on the relative qualities of the

males. Lek mating systems have been observed in at least five species of African ungulates, including the Uganda kob (*Kobus kob*) and white-eared kob (*Kobus kob leucotis*), and four species of fruit bats, including the hammerheaded bat (*Hypsignathus monstrosus*) and Wahlberg's epauletted bat (*Epomophorus wahlbergi*). These and other lek mating species occur when males are unable to economically defend either the females or the resources they require. Such a situation can arise when females exploit widely dispersed resources such as grasses (ungulates) and fruit (bats). At high female density, males of the Uganda kob form leks and display to females, but they defend harems (resource-based territories) at low density when defense is presumably economically feasible. Male hammerheaded bats display from calling roosts located in gallery forests along rivers. They emit series of sharp "honks," and when females approach, they vigorously shake their wings and increase the rate of vocalizations. Once a female chooses a particular male, they mate and the male continues to call. As in other polygynous mating systems, some males have greater success in attracting females than do others. In one study, 6% of male hammerheaded bats achieved approximately 79% of the matings in a single year. Courtship displays in species that form leks are fueled by body reserves that allow high instantaneous increases in metabolic rates (up to 20 times BMR). The short-term increase in metabolism that males incur during displays contrasts with the long-term or sustainable metabolic scope that reflects the upper limit of metabolism over a longer period of time.

Polygyny also is common among carnivores where dominant males typically roam over larger areas and seek females in estrous (e.g., felids). Although energetically costly, this strategy appears to contribute to their high reproductive success. A roaming strategy is energetically more profitable for a dominant male because the probability of taking over other females is related to his social status. When females are concentrated in patches, a dominant male can increase the number of potential matings by including a second patch within his range.

Mating systems among canids vary in relation to body size, but the pervasive mating system is monogamy. Smaller species such as red foxes (*Vulpes vulpes*) and bat-eared foxes (*Otocyon megalotis*) are usually monogamous but are sometimes polygynous. The largest canids, such as the African hunting dog (*Lycaon pictus*), are generally monogamous but with a tendency toward polyandry (several females monopolize a single male). In some situations, a

polygynous mating system may change depending on the density and dispersion of females, or it may change from polygamy to monogamy depending on the abundance of (and ability to monopolize) females. When mating is restricted to only one season of the year, the preferred resources for males may differ during the mating and nonmating seasons because the two resources (receptive females and available food) tend to have different characteristics. Where males cannot monopolize feeding, nesting, or roosting resources, monogamy becomes possible. Where habitats are characterized by having clumped resources that can be defended economically, polygyny tends to be favored.

In general, variation in mating behavior in mammals can be attributed to four characteristics of females: (1) the extent to which female reproductive rate can be increased by assistance with offspring, (2) the size of female ranges, (3) the size and stability of female groups, and (4) the density and distribution of females in space. A male is better able to monopolize several females when male care of offspring is not needed and when females have small ranges or when they occur in small stable groups. Alternatively, when females are solitary in large ranges or occur in large unstable groups, or when female density is high, it is uneconomical for males to defend several females at once.

6.2 Gamete Production

As in most other vertebrates, mammals are seasonal breeders and produce gametes only during the reproductive season. Energy is generally allocated to gonadal development prior to the breeding season in both sexes. In comparison with female mammals, males allocate a relatively small amount of energy to the production of sperm and ejaculate. The size of testes, expressed relative to body mass, is a standard method for comparing the investment among different species of mammals and birds. The variation in relative testes size in mammals ranges from 0.02 to 0.75% in primates, from 0.15 to 8.40% in rodents, and from 0.12 to 8.40% in bats. Among terrestrial mammal species that have been studied, bats have the widest range of variation in relative testes size. In fact, the relative testes size of the insectivorous Rafinesque's long-eared bat, *Corynorhinus rafinesquii*, exceeds that of all other mammal species. Still, the investment made by both sexes is a relatively trivial energy investment compared with the energy allocation to other aspects of reproduction and to the

investment that oviparous (egg-laying) vertebrates make in reproduction.

Sexual selection by sperm competition has been postulated as the mechanism that contributes to the wide range and maximal size of testes in mammals. Species in which multiple males mate with multiple females (promiscuity) have the largest testes, whereas the smallest testes occur in species where single males mate with single females (monogamy) or multiple females (polygyny). Thus, the mating system appears to be an important variable that explains testes size in mammals. In bats, sperm-storing species also have larger relative testes size than do non-sperm-storing species. Among rodents that breed two or more times per year, males allocate more resources to sperm production.

6.3 Nest Construction/Excavation and Nest Guarding

Nest construction varies widely among terrestrial mammals, with females in colder climates commonly investing more in nest construction than those in moderate to warm regions. Females of many small terrestrial mammals (e.g., shrews, rodents, lagomorphs) frequently excavate burrows or runways near the surface or create deeper cavities in the ground and construct nests made of vegetation (leaves, grass, and tree branches) where young are cared for until they are weaned. The excavated burrows and nest cavities provide protection from many would-be predators and provide insulation for the occupants, especially altricial young and their mothers. Insulation provided by nest material helps to reduce maintenance costs (for thermoregulation) of both the mother and her young, making it possible for the female to allocate more energy to reproduction. Medium-size terrestrial mammals (e.g., foxes, woodchucks, badgers) also construct burrows and nest spaces below the ground, but because they have a greater body mass and thus thermal inertia, they do not need to invest as much time and energy in actual nest construction. The Canadian beaver (*Castor canadensis*), a medium-size monogamous mammal, invests extensively in nest construction, which typically involves constructing lodges from vegetation (tree branches and grasses). Both males and females engage in nest-building activity. Before construction of a lodge, which includes the natal nest, beavers may first construct dams to increase the depth of the water so that the only access to the lodge is from an opening beneath the water. Such lodges,

with additional nest material, provide not only insulation but also protection from predators.

With the exception of large arboreal primates (e.g., chimpanzees, orangutans), most large mammals and bats do not construct nests. Instead, large mammals seek shelter from inclement weather by forming tightly spaced social groups (e.g., musk oxen, elephants) or seek shelter beneath trees (e.g., many bovids and cervids) or under rock overhangs in treeless habitats (e.g., asses, mountain sheep). Because most large mammals produce precocial offspring, extensive natal nests are not needed. Although bat species are altricial at birth, the parents do not construct shelters or nests but rather use existing structures such as tree cavities, caves, mines, buildings and similar spaces, and roost beneath foliage. In contrast to some mammals (mostly rodents) that seek shelter in similar places and use nesting material from grass and leaves, bats do not use any such material. Sites selected by female bats for rearing young are often physically confined spaces (cavities in trees and caves) that are warmer than alternative locations and, thus, provide conditions that reduce maintenance costs and promote milk production by females and growth of their young. Because females often return to these same roosts year after year, little energy is invested in finding these sites. Energetic benefits accrue to both the mothers and their young by occupying shelters in which metabolic heat can be trapped, thereby reducing heat loss to an otherwise cool environment. However, a few species of bats construct shelters by modifying leaves, other plant parts (roosts and vines), bird nests, and nests of ants and termites that, in turn, often become occupied by females to give birth to young. These so-called "tents" are constructed by single males that recruit females with whom they have a high probability of mating. Because these shelters are constructed by males that are chosen by females to rear their young, tent construction can be considered as a form of paternal investment.

Natal nests of many mammals are often strongly defended from predators and conspecific intruders by both males and females. Many colonial species (e.g., prairie dogs, marmots, bats) use a variety of signals to alert nest or group members to threats from such predators, and these may include vocalizations (e.g., chirps, howls), visual signals (e.g., flashes of white or black tails by prairie dogs, wing displays by bats), and olfactory signals (e.g., production of scent from specialized glands, secretion into the urine). Some rodents secrete substances into their urine when they are stressed, and this acts as a pheromone that may cause others to escape or avoid an area. Skunks and

other mustelids are well known for the use of visual and scent displays that they use when threatened by conspecifics and would-be predators.

6.4 Gestation and Size at Birth

Gestation in mammals occurs in the uterus; however, the development of the uterus and its role in gestation differs considerably among different mammalian taxa. In placental (eutherian) mammals, an intimate physical, nutritive, and endocrine association develops between the fetus and the uterus via a chorioallantoic placenta, which nourishes the fetus for the duration of gestation and helps to regulate the duration of gestation. In contrast, the uterus of marsupial (metatherian) mammals and associated choriovitelline (yolk sac) placenta plays a relatively minor role in embryonic development. No intimate physical or endocrine association is established, and the young are born in an extremely altricial state. However, in either case, the requirements for gestation depend on the amount of energy and matter retained by the gravid uterus and the enlarging mammary glands and the increased maintenance costs of these structures. Altricial marsupials are instead nourished throughout most of their "fetal" period from highly specialized mammary glands and nipples.

Embryonic development incurs the greatest cost to the female during gestation. In eutherian mammals, fetal growth is curvilinear, with most of the increase in embryonic mass occurring after 50 to 60% of the gestation period has elapsed. In general, the length of the gestation period and the birth mass of eutherian neonates increase as a power of maternal mass, with smaller species having shorter gestation periods than larger ones and with smaller species having smaller offspring than larger ones. However, carnivores, primates, and most bats have longer gestation periods than can be predicted from maternal body mass, whereas marsupials and red and giant pandas have shorter gestation periods than can be predicted from maternal mass. Exceptions to predictions in length of gestation occur predominantly among marsupials that have a choriovitelline placenta and have unusually short gestation periods and among eutherian mammals that undergo delayed embryonic development (e.g., mustelids, pinnipeds, some bats). For example, the gestation length of an 18-g rodent is approximately 21 days, whereas the gestation period of comparable size bat species can range upward from 2 to 6 months. Bats and bears best exemplify exceptions to the predicted trend in size at birth

among eutherian mammals. Newborn little brown bats (*Myotis lucifugus*) are larger at birth than predicted, weighing about 2 g and representing nearly one-fourth of the mother's body mass. By contrast, newborn black bears (*Ursus americanus*) are smaller than predicted, ranging in mass from 200 to 450 g and averaging 1/250th of the mother's mass.

In general, birth mass is a reflection of the metabolic body mass of the mother, with larger species producing offspring that represent a smaller proportion of adult mass. The energy cost of gestation can be estimated using birth mass to maternal mass relationships, a maternal mass to gestation period regression, the composition and growth characteristics of the gravid uterus, and an estimate of the oxygen consumption of the gravid uterus per kilogram of tissue. Several studies have shown that the maintenance cost of the gravid uterus is the major cost of gestation in eutherian mammals. In general, gestation is far less costly and stressful in larger mammals than in smaller mammals because the relatively smaller neonate of larger species is produced over a much longer period. However, when one compares the total cost (maintenance and production) of a gravid uterus with that of the gestation period, the cost is proportional to maternal metabolic body mass ($M^{0.77}$). In eutherians, protein accumulates curvilinearly in the gravid uterus and mammary glands during gestation. In general, because of longer gestation periods in larger species, protein accumulates at a decreasing rate per kilogram of fetus. Changes in fetal size of gestation length are far more complex than simply comparing relative growth rates. Pregnant females can benefit by lengthening the gestation period if protein (or minerals) is limiting, but they gain nothing by prolonging the gestation period if energy is limiting. Some mammals ingest nutrients in excess of the actual costs of gestation, and because lactation is more costly than gestation, females of larger species (but not small species) can accumulate reserves during gestation when food is available and can mobilize it later for milk synthesis. Mammary glands begin to enlarge during the final trimester of gestation and, thus, can be considered as a cost of pregnancy and lactation.

6.5 Lactation

The production and secretion of milk is a trait that distinguishes mammals from all other vertebrates, although some birds (e.g., pigeons, doves, penguins, flamingos) feed crop secretions, or "crop milk," to

hatchlings during their first few days of life. Among mammals, the milk produced by a postpartum female provides an important dietary buffer between the dependent fetus and the weaned and nutritionally dependent young. Milk is produced by mammary glands that vary in position from the thoracic to the inguinal region of the body. Mammary glands generally vary from a single pectoral pair in primates, elephants, and bats up to 12 pairs located in the abdominal and inguinal regions in some rodents and marsupials. The young of placental and marsupial mammals obtain milk from nipples by a sucking action that stimulates milk letdown (the neural-hormonal release of milk from the mammary glands). Nipples are absent in monotremes (egg-laying mammals), and milk is secreted from mammary glands directly onto the ventral surface of the skin of females so that the beaked young can lap up the milk with their tongues.

Female mammals typically provision their offspring with milk for varying amounts of time. Factors known to influence the length of lactation is adult body size relative maturity of their young at birth (altricial or precocial) as well as the quantity and quality of milk. The quantity and quality (energy and nutrient content) of milk can vary according to a particular species, the type of food that is ingested by the mother, the availability of water, and the type of habitat in which they live. Terrestrial mammals that have a high-fat diet (those that live in arid regions) often produce milk with high lipid content. In contrast, mammals that feed on fruit and nectar or fleshy aquatic vegetation generally produce milk that is high in water content and carbohydrates but low in protein. Carnivorous and insectivorous species more commonly produce milk that is high in protein and fat but low in carbohydrates. The milk composition may also change over the course of lactation in many mammals. The first milk produced during a lactation cycle is called colostrum. Colostrum has a high concentration of antibodies and active phagocytic cells, each of which are thought to be important for the development of an effective immune system in neonates.

The major nutrient constituents of milk are water, minerals, protein (casein and whey), fat, and carbohydrates. Protein concentration ranges from less than 2% in primates to nearly 16% in the eastern cottontail (*Sylvilagus floridanus*). Fat content in the milk of terrestrial mammals also varies among species, ranging from a mean of 2.3% in lemurs (*Lemur* sp.) to 18.5% in the brown bear (*Ursus arctos*). The milk composition of bats also appears to

be influenced by diet. For example, the milk composition of insectivorous species generally has a greater amount of dry matter, fat, and protein content than that of omnivorous and frugivorous species. The relatively low fat, protein, and dry matter content of milk of frugivorous species is consistent with the relatively low fat and protein content of available fruit. Insectivorous species produce milk with percentages of fat and protein that often exceed values reported for other small mammals, an observation that is consistent with evidence that small mammals produce milk of higher energy content than do most large mammals.

Some females suckle their young on demand (e.g., marsupials, primates, some ungulates, some bats), whereas others suckle their young on a schedule (e.g., lagomorphs, most rodents, most carnivores, many artiodactyls, some bats). As their young increase in size, so do the energy and nutrient demands on the mothers to produce more milk. Thus, the composition of milk often changes over the course of lactation. For example, in the Brazilian free-tailed bat (*Tadarida brasiliensis*), fat content of milk changes from a mean of 17.3 to 25.8% from early to peak lactation. Large mammals such as carnivores, bovids, and cervids also deposit fat reserves during pregnancy and mobilize them during lactation to either support the production of milk or support the increased demands for energy of the mother to secure food. During peak lactation, when energy demands are the highest, females may mobilize their other body reserves to supplement daily food intake. This is especially the case during periods of food shortages but also in small mammals that do not have the capacity to store large fat reserves.

As energy demands increase due to lactation, maintenance costs also increase because more time is invested in locomotor activity to secure food. The amount of time and energy that mammals allocate to foraging can increase more than twofold compared with that allocated to pregnancy. For example, female little brown bats (*M. lucifugus*) and Brazilian free-tailed bats (*T. brasiliensis*) more than double their nightly intake of insects from early to peak lactation. As lactation progresses, the energy cost of milk synthesis increases, corresponding to the increased energy demands of the growing young. Milk production generally increases until peak lactation and then begins to decrease as infants become more independent and begin to feed on other foods.

Estimates of milk intake by maternally raised young, when combined with estimates of milk composition, provide data necessary for determining

requirements for maternal production and requirements for neonatal growth. Methods used for estimating milk intake by suckling young include weighing the neonate before and after a controlled nursing period, hand milking a lactating female after a controlled period of neonate removal, and using the dilution of isotopes (usually deuterium or tritium oxide) to quantify the dilution of the body water pool by ingested milk. The isotope dilution method offers the only available method for estimating milk intake in free-ranging mammals. Thus, milk intake can be directly estimated from measurements of neonatal water kinetics. Because milk is the only source of water to a suckling before it takes solid food, measuring the daily water flux of such a neonate using isotope dilution and estimating the proximate composition of milk (energy content of the dry matter) makes it possible to estimate the amount of energy that a mother invests in her offspring in a given 24-h period or over the course of lactation. The energy expenditure of free-ranging female mammals during lactation can be quantified using a combination of time budget analysis and laboratory respirometry or from estimates of energy expenditure using the doubly labeled water technique.

Energy requirements for pregnancy and lactation are proportional to the productive process involved—synthesis of fetal, placental, uterine, and mammary tissue during pregnancy and milk production during lactation. In one study, metabolizable energy requirements for domestic ruminants and horses increased by 28.2 and 23.2% above maintenance levels, respectively, during late pregnancy and by 37.2 and 41.5% from late pregnancy to early lactation. In free-ranging golden mantled ground squirrels (*Spermophilus lateralis*), the amount of metabolized energy allocated to reproduction by females into newborn young and milk export increased by 17.7% over nonreproductive levels in one study. The total energy budget of a lactating female includes loss or storage of body reserves and the amount of energy that is produced and exported as milk (Fig. 1). In free-ranging little brown bats (*M. lucifugus*), average daily energy requirements during lactation (litter size of one) exceeded requirements during late pregnancy by a factor of 1.2 in one study. However, when milk output was included, energy required by females during lactation nearly doubled from early (33.8 kJ/day) to peak lactation (60.3 kJ/day).

Among terrestrial mammals, most young are weaned by the time they have reached approximately 40% of adult size. In contrast, lactating female bats typically nurse their young until they have nearly

reached adult size. Because bats cannot fly until they are nearly adult size, mothers must continue to suckle their young until such time that they have achieved approximately 90% of adult wing dimensions and can secure food on the wing. The lactation period for most small (<60 g) insectivorous bats is 6 to 8 weeks. Larger mammals (>10 kg) have disproportionately longer lactation periods (several months) and distribute their total costs over a far greater period of time. Because larger species tend to produce relatively low levels of milk at peak lactation and during the prolonged end of the lactation cycle, the average daily cost as a function of BMR is lower in larger species. Thus, lactation costs are high but are of relatively short duration in small mammals, as compared with a relatively lower cost but longer duration in large mammals.

Because most mammals are polygynous, males assume little or no role in the care of the young they have fathered. However, among monogamous species, females suckle their young but may be assisted by fathers who provision their young with food (e.g., many carnivores). In general, male mammals do not lactate; however, recent observations have shown that some male mammals do produce milk, perhaps as a trait in some monogamous species or possibly as a consequence of eating leaves high in phytoestrogens or eating food that has been exposed to pesticides that have estrogenic properties. Males of at least two species of Old World fruit bats (*Dyacopterus spadiceus* and *Pteropus capistratus*) are known to produce milk, although it has not been determined that they actually suckle their offspring. Because the mating system of these two species is not known, any relationship between a monogamous mating system and male lactation remains unclear.

In addition to productive processes (fetal growth and lactation) associated with pregnancy and lactation in mammals, other energy costs include the establishment and maintenance of territories associated with courtship and mating, gamete production (relatively small in mammals), the increased locomotor activity associated with food acquisition, nest construction and guarding, and other forms of parental care.

6.6 Parental Care

Because of constraints imposed by polygynous mating systems and female lactation, parental care in mammals generally is largely the responsibility of females. However, in some species that breed cooperatively, offspring from previous litters help

care for members of the subsequent litter. Events and activities associated with parental care in mammals cover gestation, parturition (birth), lactation, grooming, feeding, and tutoring of young. During or before the gestation period, nest construction and nest defense are the primary energy-demanding activities that extend above maintenance costs. Parturition is an energy-demanding activity, but because this period is relatively short (minutes to hours), the overall impact on the energy budget of a female is trivial. During lactation, thermoregulation and grooming are the primary activities. Among these, lactation is the most energetically demanding. The energy demands associated with lactation will depend on the duration of lactation, the number and size of offspring, the body size of the mother, the local microclimate, and the availability of food resources. For species in which there is a prolonged postnatal period of growth and development (e.g., primates, elephants), tutoring exacts additional energy costs.

Among monogamous mammals, males invest in their offspring both indirectly and directly. For example, a dominant male gray wolf (*Canis lupus*) or dominant male coyote (*Canis latrans*) will bring food to a lactating female in its den, and when the young approach the age of weaning, he will offer them partially chewed food. Males may defend burrows that contain neonates from potential predators, either by giving alarm calls, signaling with visual displays, or distracting potential predators away from nest area. Paternal care that is most often associated with monogamy is relatively rare in placental mammals, averaging less than 5% of all species that have been studied.

Alloparental care has also been observed in several species of mammals. Among these are highly social and cooperative breeding species. For example, alpine marmots (*Marmota marmota*) live in colonies that include a territorial pair, subordinates, yearlings, and young of the year. Subordinates cluster with newborns, groom them, and cover them with grass, behaviors that improve the survival of offspring. A special case of alloparental behavior has been observed in eucocial mammals, including the naked mole rat (*Heterocephalus glaber*) and Damara mole rat (*Cryptomys damarensis*). Colonies of naked mole rats consist of one reproductive female, several males; and several nonreproductive female workers. Reproductive maturity of subordinate females is suppressed pheromonally by the presence of a dominant "queen" who mates with and gives birth to offspring fathered by a few resident males. The reproductively suppressed females construct burrows and nest chambers

and assist the queen by bringing her food and grooming her. The dominant female suckles all of the offspring. Alloparental behavior also extends to assistance during the birthing period. Female Rodrigues fruit bats (*Pteropus rodrigues*) have been observed assisting other females during the birthing period. In one instance, when a female was having a difficult (breech birth), another female approached the expecting mother and groomed her pup as it was emerging from the vagina. She also tutored the mother in the correct birthing position and wrapped her wings around the mother and pup during the birth process. Similar observations of helper-assisted births have been observed in the common marmoset (*Callithrix jacchus*), African elephant (*Elephas maximus*), Afghan hunting dog (*Canis familiaris*), and spiny mouse (*Acomys cahirinus*). Sharing of a blood meal with roostmates, including young, has been observed in the common vampire bat (*Desmodus rotundus*). This behavior appears to be practiced mostly by mothers who provision their own young, but it has also been observed between adults as well as between adults and related but nonfilial young. This altruistic behavior appears to be based on reciprocity and kin selection in this highly social and communally breeding species. Parental care in mammals, especially in highly social species (e.g., primates, some rodents, bats), also involves social grooming and protecting young from extremes in temperature or inclement weather. The latter may also involve clustering and huddling to reduce energy expenditure.

7. CONCLUDING REMARKS

This article has provided a brief overview and a few glimpses of how various factors influence the reproductive energetics of terrestrial reptiles, birds, and mammals. The amount of energy allocated to various phases of reproduction (courtship and mating, incubation, gestation, lactation, parental care, and the pre- and postfledging or postlactational periods) is, in part, influenced by the type of reproduction (oviparity vs. viviparity), different types of mating system (monogamy, polygyny, and polyandry), rates of posthatching and postnatal growth, competitive interactions with members of the same or different species, and extrinsic factors such as habitat quality, food availability, and climate. In general, costs of courtship and mating are higher for males that typically compete for females by developing elaborate structures or behaviors used in

courtship display to show females that they are fit and by defending either mating territories or females. The energy investment that male vertebrates make in the production of gametes is relatively small compared with that made by females, and female reptiles and birds typically invest more in egg production (both size and number) than do female mammals.

Apart from the production of a relatively large number of eggs, few reptiles exhibit maternal care once the young hatch. In contrast, birds and mammals often invest extensively in their young once they hatch or are born. Female birds allocate large amounts of energy to brooding and feeding their young until they have reached or exceed adult size. Some species invest further in their offspring by tutoring them once they have fledged. In contrast, mammals typically produce a relatively small number of eggs that are fertilized internally, the fetus develops in the uterus that is protected from environmental extremes, young are born alive, and females nurse their young with milk for variable periods of time and in many species may provide extended care to their offspring. Among males, courtship and mating are the most energetically expensive activities in reptiles, birds, and mammals. The production of nutrient- and energy-rich eggs is the most energetically expensive phase of reproduction in female reptiles. Production and incubation of eggs and feeding of nestlings comprise the most energetically expensive phase of reproduction in female birds. Lactation is the most energetically expensive period of reproduction in female mammals.

SEE ALSO THE FOLLOWING ARTICLES

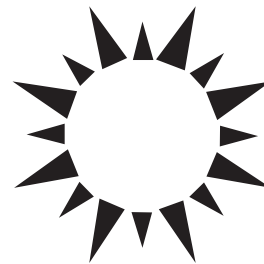
Conversion of Energy: People and Animals • Food Capture, Energy Costs of • Human Energetics • Migration, Energy Costs of • Thermoregulation

Further Reading

- Blem, C. R. (2000). Energy balance. In "Sturkie's Avian Physiology" (G. C. Whittow, Ed.), 5th ed., pp. 327–341. Academic Press, San Diego.
- Bradbury, J. W., and Vehrencamp, S. L. (1998). "Principles of Animal Communication." Sinauer Associates, Sunderland, MA.
- Carey, C. (ed.). (1996). "Avian Energetics and Nutritional Ecology." Chapman & Hall, London.
- Clutton-Brock, T. H. (1991). "The Evolution of Parental Care." Princeton University Press, Princeton, NJ.
- Clutton-Brock, T. H., Guinness, F. E., and Albon, S. D. (1982). "Red Deer: Behavior and Ecology of Two Sexes." University of Chicago Press, Chicago.
- Congdon, J. D., Dunham, A. E., and Tinkle, D. W. (1982). Energy budgets and life histories of reptiles. In "Biology of the Reptilia" (C. Gans and F. H. Pough, Eds.), Vol. 13, pp. 233–271. Academic Press, New York.
- Cummins, J. M., and Woodall, P. F. (1985). On mammalian sperm dimensions. *J. Reprod. Fertil.* 75, 153–175.
- Eisenberg, J. F. (1981). "Mammalian Radiations." University of Chicago Press, Chicago.
- Feldhamer, G. A., Drickamer, L. C., Vessey, S. H., and Merritt, J. F. (2003). "Mammalogy: Adaptation, Diversity, Ecology." McGraw-Hill, New York.
- Gittleman, J. L. (ed.). (1989). "Carnivore Behavior, Ecology, and Evolution." Cornell University Press, Ithaca, NY.
- Gittleman, J. L., and Thompson, S. D. (1988). Energy allocation in mammalian reproduction. *Am. Zool.* 28, 863–875.
- Hosken, D. J. (1997). Sperm competition in bats. *Proc. Roy. Society London B* 264, 385–392.
- Kunz, T. H., and Fenton, M. B. (eds.). (2003). "Bat Ecology." University of Chicago Press, Chicago.
- Kunz, T. H., and Hood, W. H. (2000). Parental care and postnatal growth in the *Chiroptera*. In "Reproductive Biology of Bats" (E. G. Crichton and P. H. Krutzsch, Eds.), pp. 415–468. Academic Press, San Diego.
- Kunz, T. H., and Lumsden, L. F. (2003). Ecology of cavity and foliage roosting bats. In "Bat Ecology" (T. H. Kunz and M. B. Fenton, Eds.), pp. 3–89. University of Chicago Press, Chicago.
- Kunz, T. H., and Nagy, K. A. (1988). Methods of energy budget analysis. In "Ecological and Behavioral Methods for the Study of Bats," (T. H. Kunz, Ed.), pp. 277–302. Smithsonian Institution, Washington, DC.
- Lacey, E. A., Patton, J. L., and Cameron, G. N. (eds.). (2000). "Life Underground: The Biology of Subterranean Rodents." University of Chicago Press, Chicago.
- Lacey, E. A., and Wiczorek, J. R. (2001). Territoriality and male reproductive success in arctic ground squirrels. *Behav. Ecol.* 12, 626–632.
- Loudon, A. S. I., and Racey, P. A. (eds.). (1987). "Reproductive Energetics in Mammals." Oxford University Press, Oxford, UK.
- McCracken, G. F., and Wilkinson, G. S. (2000). Bat mating systems. In "Reproductive Biology of Bats" (E. G. Crichton and P. H. Krutzsch, Eds.), pp. 321–362. Academic Press, San Diego.
- McNab, B. K. (2002). "The Physiological Ecology of Vertebrates: A View from Energetics." Cornell University Press, Ithaca, NY.
- Nagy, K. A. (2001). Food requirements of wild animals: Predictive Equations for free-living mammals, reptiles, and birds. *Nutr. Abstr. Rev.* 71, 21R–32R.
- Oftedal, O. T. (1985). Pregnancy and lactation. In "Bioenergetics of Wild Herbivores" (R. J. Hudson and R. G. White, Eds.), pp. 215–238. CRC Press, Boca Raton, FL.
- Orrell, K. S., Congdon, J. D., Jenssen, T. A., Michener, R. H., and Kunz, T. H. (2003). Comparison of energy expenditure by male and female *Anolis carolinensis* during breeding and postbreeding seasons. *Physiol. Biochem. Zool.*, in press.
- Orrell, K. S., Congdon, J. D., Jenssen, T. A., Michener, R. H., and Kunz, T. H. (2004). Intersexual differences in energy expenditure of *Anolis carolinensis* lizards during breeding and postbreeding seasons. *Physiol. Biochem. Zool.* 77. In press.
- Peaker, M., Vernon, R. G., and Knight, C. H. (eds.). (1984). "Physiological Strategies in Lactation." Symposia of the Zoological Society of London No. 51 Academic Press, London.

- Ricklefs, R. E. (1974). Energetics of reproduction in birds. *In* "Avian Energetics" (R. A. Paynter, Jr., Ed.), pp. 152–297. Nuttall Ornithological Club, Cambridge, MA.
- Robbins, C. T. (1993). "Wildlife Feeding and Nutrition." Academic Press, San Diego.
- Stearns, S. (1992). "The Evolution of Life Histories." Oxford University Press, New York.
- Voigt, C. C., von Helversen, O., Michener, R., and Kunz, T. H. (2001). The economics of harem maintenance in the sacwinged bat, *Saccopteryx bilineata*. *Behav. Ecol. Sociobiol.* 50, 31–36.
- Wilkinson, G. S., and McCracken, G. F. (2003). Bats and balls: Sexual selection and sperm competition in the *chiroptera*. *In* "Bat Ecology" (T. H. Kunz and M. B. Fenton, Eds.), pp. 128–155. University of Chicago Press, Chicago.

Research and Development Trends for Energy



PAUL J. RUNCI and JAMES J. DOOLEY

Joint Global Change Research Institute
College Park, Maryland, United States

-
1. The Importance of Energy R&D
 2. Energy R&D Performers and Sponsors
 3. Motivations for Energy R&D
 4. Trends in Energy R&D Investments
 5. Assessing R&D Trends: Methodological Challenges
 6. Conclusion

Glossary

development A stage of the innovation process during which basic research findings are applied to the development of new materials, devices, or processes for practical use.

energy restructuring Often referred to as energy deregulation or liberalization, the process of opening the energy industries (e.g., electricity, gas, oil) to market forces and competition; in many countries, the process involves the dismantling of large, integrated state monopoly energy companies and the privatization of all, or of large parts of, the industries.

energy technologies Materials, devices, and processes for the location, recovery, preparation, storage, transport, conversion, and use of energy resources and for the treatment of associated waste products.

innovation The systematic practice of research and development and the creation of new technologies, usually through a linked network of scientists, engineers, and research sponsors at university, private sector, and government-owned research facilities.

research The initial stage of the innovation process where new knowledge of the natural world is sought through experiment and observation without specific motivation for the application of findings to particular practical problems or to the creation of new technologies.

Energy research and development (R&D) is often defined in a manner similar to that in which a U.S. Supreme Court justice defined pornography: "I know

it when I see it." That is, most of those concerned with energy R&D believe that there is a common tacit definition, yet attempts at explicit definition usually reveal significant differences in understanding among researchers, analysts, and those who support energy innovation. Accepting that there are few explicit definitions of energy R&D provided in the literature, and that no extant definition is regarded as constituting a standard one, this article adopts the following definition based on that of the U.S. President's Committee of Advisors on Science and Technology:

Energy R&D refers to a series of linked processes by which technologies for energy supply, end use, or carbon management move from theoretical conceptualization to feasibility testing and, ultimately, to small-scale deployment. Energy R&D encompasses both basic and applied research, technology development, and demonstration associated with each phase of the energy life cycle, including production (e.g., mining, drilling), energy conversion and power generation (e.g., nuclear fission and fusion, fossil and renewable energy systems, bioenergy, hydrogen production), transmission, distribution, energy storage, end use and energy efficiency, and carbon management. Carbon management technologies aim to manage anthropogenic releases of greenhouse gases, such as those associated with the combustion of fossil fuel use, in an effort to mitigate the potential impacts of these emissions on climate systems. Carbon management technologies include advanced agroforestry practices aiming to enhance the capacity of soils, trees, and other standing biomass to absorb atmospheric carbon dioxide; engineered pre- and postcombustion carbon capture systems; and technologies for geologic and ocean carbon sequestration.

This definition is an extension of one first put forward by the U.S. President's Council of Advisors on Science and Technology in 1999. Although this constitutes a useful and comprehensive working definition of energy R&D, it is important to note that it is descriptive rather than prescriptive in

nature; no government agency or private organization uses such a definition as a guide in planning its energy R&D investments. Although many commercially available technologies, such as low-emissivity windows, are the result of established energy R&D programs, other projects and technologies are often classified as the fruit of energy R&D investments *ex post facto*. For example, the use of natural gas turbines for electric power production has been one of the most important energy technology advances of the past few decades, yet the investments that led to this breakthrough were largely made by the U.S. Department of Defense during the 1960s and 1970s under the rubric of aircraft engine research. That this work led in the long run to the technology's application in the electric power industry raises taxonomical questions about its retroactive classification as "energy R&D." Thus, defining energy R&D remains a malleable enterprise.

1. THE IMPORTANCE OF ENERGY R&D

Energy technologies are the primary determinants of energy availability, fuel choice, end use efficiency, and the degree and nature of by-product emissions, and energy research and development (R&D) is the vehicle by which new technologies become available. To the extent that energy technologies augment or diminish society's range of energy choices, energy R&D is a matter of high importance. Over the past half-century or so, awareness of energy use and its implications for economic growth, national security, and environmental quality have at times held prominence to an unprecedented degree. To a large degree, these three societal goals have formed the underlying rationale for investments in energy R&D across the globe over this period. The relative emphasis placed on each of these goals has varied from country to country and even within countries over time.

For example, two of the three founding treaties of the European Union focused on the coordinated management of energy resources and energy R&D for reconstruction and economic growth. Moreover, virtually all industrialized countries created domestic agencies for energy technology development during the post-World War II period. For energy resource-poor countries such as Japan, the fact that energy R&D continually receives a sizable percentage of government's total outlays for all kinds of R&D (more than 10% of all government R&D expendi-

tures consistently over the past 25 years or so) is a testament to the importance of this activity. At other times and in countries such as the United States, Canada, and most of the European Union, the importance of energy and energy R&D has received relatively little attention, and as a result, public and private sector investments in energy R&D have suffered. Energy R&D tends to be a near- to midterm beneficiary of energy crises.

Energy considerations are central factors in several policy areas, including environmental, economic, and defense policy. Energy resources have served both as a cause for and as a weapon of war, and the consequences of energy use have manifest themselves in a host of environmental problems, ranging from local air pollution to global climate change. Consequently, energy R&D and the technologies that the R&D process yields broaden or narrow the portfolio of choices available to policymakers on a number of important issues.

2. ENERGY R&D PERFORMERS AND SPONSORS

Energy R&D has historically been the province of large industrialized countries with limited domestic energy resources and high levels of energy imports. Most of the world's energy R&D is funded and performed by a small group of industrialized countries: the United States, Canada, Japan, Germany, France, Italy, The Netherlands, Switzerland, the United Kingdom, and countries comprising the European Union. According to the International Energy Agency (IEA), these countries alone account for more than 90% of the industrialized world's energy R&D. Although these countries are clearly the "energy R&D-intensive" leaders, only in the case of Japan does energy R&D account for more than 1.5% of a nation's overall R&D enterprise.

The performance of energy R&D in the private sector is similarly concentrated in large industrial firms. For example, in the United States in 2000, firms that had more than 10,000 employees performed 89% of industrial energy R&D. Although it is clear that these companies are the leading performers of energy R&D, it is also clear that the performance of R&D is not a significant aspect of their business. The energy industries are among the least R&D intensive in the industrialized world. In comparison, the most R&D-intensive industries, such as medical equipment and information technology, invest 12 and 8% of annual

sales, respectively, in R&D. Although a few specialized oil and gas exploration service companies have R&D/sales ratios in excess of 7%, this level of investment is the exception within the energy industries that typically invest less than 1% of sales in R&D.

Neither of these observations is meant to exclude the possibility (indeed the certainty) that smaller countries and smaller firms are not conducting important energy R&D. Rather, it seems clear that the performance of energy R&D is limited to this select set of large industrialized countries and large firms by the often times significant research infrastructure requirements. (It is hard to imagine a very small firm conducting all of the research needed to bring about a new fission reactor from conception to pilot scale testing.) The successful performance of energy R&D often depends on the existence of a large cadre of scientists and engineers, the availability of specialized laboratories and user facilities, funding mechanisms for sustained research support, and programs for technology commercialization and adoption. A relatively small set of countries and large firms have these capabilities; therefore, it is these entities that constitute the bulk of the global energy R&D enterprise.

Governments have historically been the principal sponsors of energy R&D, whereas the private sector has played an important but secondary role from the perspective of overall investment levels. Governments have historically been the main supporters of long-term, higher risk R&D in particular, whereas private sector investments have focused more on short-term research and technology commercialization efforts. In several industrialized countries, including the United States, Germany, and Japan, the gap between public and private sector investments has widened over the past 15 years or so. For example, whereas government and industry investments were roughly equal in 1990, government now invests approximately 50% more than the private sector in the United States. Similarly, whereas the German government invested 30% more than did industry in 1990, by 2000 its energy R&D investment was more than double that of the private sector.

3. MOTIVATIONS FOR ENERGY R&D

As noted previously, in most countries, public sector energy R&D has been closely tied to domestic policy priorities such as efforts to bolster national security, to enhance economic productivity, and to remedy

specific local- to regional-scale environmental problems. Because policy priorities change continually, so too have the focus and intensity of R&D efforts in performing countries over the past half-century. For example, the oil supply shocks of the 1970s and the related economic crises were among the strongest drivers of energy R&D efforts in industrialized countries. Between 1973 and 1981, government investments in energy R&D rose to their highest levels as new programs and new institutions formed in response to oil supply interruptions initiated by the Organization of Petroleum Exporting Countries (OPEC). Domestic policy programs, such as Project Independence in the United States, and agencies with significant energy R&D missions, such as the U.S. Department of Energy (DOE), Britain's Department of Energy, and the Organization for Economic Cooperation and Development's (OECD) IEA, grew out of Western policymakers' ambitions to reduce exposure to future energy supply interruptions. Nuclear fission and fusion research programs, energy efficiency, and renewable energy research all received substantial funding increases in the interest of reducing the energy security risks associated with high levels of fossil fuel import dependence. In real inflation-adjusted terms, energy R&D investments hit their high-water marks during this period and have been on a steady decline, for the most part, since the late 1970s and early 1980s. For example, U.S. energy R&D expenditures in 1980 were nearly three times what they are at the turn of the 21st century.

These investments in energy R&D have yielded significant economic benefits. For example, over the 20-period between 1973 and 1993, energy productivity—the amount of energy required to produce \$1 of gross domestic product (GDP)—rose by 1.7% annually on average in the United States and by as much as twice that rate in Japan and Europe. The decoupling of energy consumption from economic growth also dismissed a widely held belief regarding the tight interdependence of these two factors. In addition, overall levels of energy import dependence in industrialized countries declined for more than a decade, although they have since rebounded to historically high levels.

Although improvements in energy productivity over this period cannot be attributed exclusively to investments in energy R&D, these investments were an important factor. As the National Research Council found in its 2001 study of DOE-sponsored energy efficiency and fossil fuel R&D programs, real economic benefits were realized from DOE and private sector investments between 1978 and 2000.

Although other factors, including structural changes in the U.S. economy, fuel prices, and changes in the fuel mix, also played important roles, the National Research Council panel found that a small group of government programs exercised tremendous leverage in the U.S. economy. For example, government-sponsored programs addressing building efficiency and fossil fuel-related NO_x emissions realized benefits in the form of avoided energy and environmental costs that were orders of magnitude larger than the associated R&D investments.

In the private sector, profitability and the need to meet ever tighter environmental regulations have been the prime motivators for investment in energy technology development. Private energy R&D efforts intensified during the 1970s and 1980s in response to the higher market prices that made energy production more profitable for those firms that were energy producers and provided a significant motivation for energy users to find ways in which to conserve energy as a cost-cutting measure. For example, for oil and gas companies, the prospect of higher profits prompted major new investments in exploration and production technology aimed at bringing to market energy resources that had been considered uneconomical at lower prices. Important technological advances, such as three-dimensional seismic imaging and directional drilling, occurred during this period and have continued to buoy the profitability of energy companies through subsequent and prolonged periods of low energy prices.

For gas and electric utilities in many countries, the motivation for energy R&D stemmed from energy providers' "regulatory compact" with government. Typically, regulated utilities were guaranteed a geographically isolated market and a fixed rate of return on their capital and R&D investments, minimizing the risks associated with those investments. Moreover, in some cases (e.g., the United States), utilities were required to invest a specified portion of annual revenues in common associations that served as industry-wide R&D pools. The Electric Power Research Institute and the Gas Research Institute derived their existence largely from the mandated "wires and pipeline charges" paid by power and gas suppliers in the regulated energy marketplace. However, the process of energy deregulation has usually included the repeal of government-mandated R&D contributions. As a result, collective R&D institutions have seen their annual revenues fall by more than 50%.

The deregulation of the energy industries raises lingering questions about the current and future

motivations for energy R&D investments. Although major problems such as climate change and energy security are indicative of the significant energy challenges that lie ahead, the structure of market and policy incentives have shifted in ways that motivate public and private sector decision makers to forgo higher overall levels of energy R&D investment. Arguably, the most pressing problem facing energy R&D sponsors and performers today is the sharp dislocation between current incentives for energy R&D investment and the likely future societal demands for new generations of cleaner and more efficient energy technologies.

4. TRENDS IN ENERGY R&D INVESTMENTS

According to the IEA, the industrialized world's overall investment in energy R&D declined by approximately 17% during the 1990s (Fig. 1). The IEA's 1999 aggregate estimate of world energy R&D investment, \$7.4 billion, reflects a decrease of 50% from the high-water mark of \$14.7 billion in 1980.

Of the group of nine countries that constitute the world's leading energy R&D sponsors (the United States, Japan, France, Germany, Italy, Canada, The Netherlands, Switzerland, and the United Kingdom), seven have reduced their expenditures over the past decade or more. In each of these countries, overall R&D spending has increased significantly in real terms. Some areas, such as health and biomedical research and information technology, have experienced large and sustained R&D investment increases, reflecting the major advances in these fields over the past two decades or so and the widely held belief that the private and public returns on additional investments in these areas will be significant. The absence of acute energy crises and perceptions of shortage, such as those that characterized the 1970s and early 1980s, have discounted the value of government-sponsored energy R&D in the eyes of policymakers and the general public. Thus, as a research priority, energy R&D has lost ground in both relative and absolute terms as an element of the broader research portfolios of most industrialized countries. Only France and Japan have increased their energy R&D investments in real terms since 1991 (Table I).

Although the aggregate trend data reveal steadily waning investments, individual energy technology areas are flourishing in some countries. For example, because hydrogen and fuel cells are widely regarded

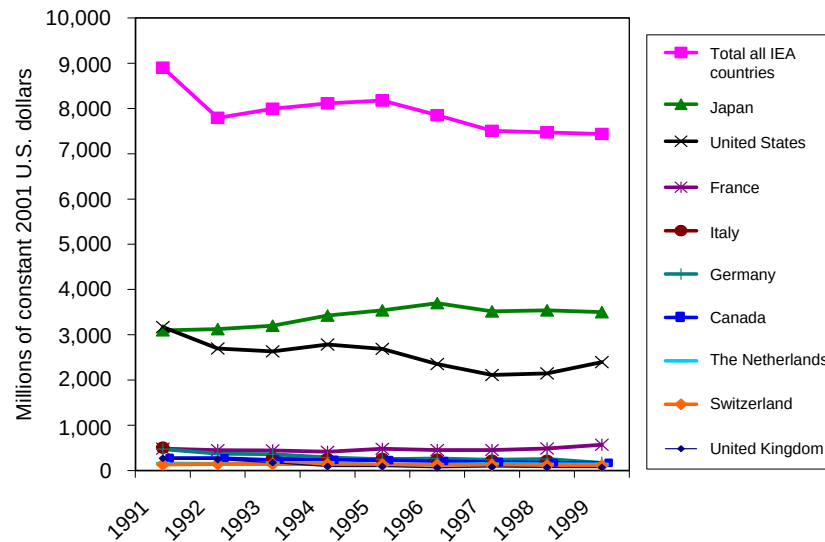


FIGURE 1 Energy R&D investment trends in major performing countries during the 1990s, as reported to the International Energy Agency (IEA). Note that currency conversions in the figure are based on market exchange rates rather than on purchasing power parities. Because energy R&D is largely a nontraded good (i.e., funded and produced within national borders rather than across them), the use of market exchange rates in converting local currencies to constant U.S. dollars may distort the real values of countries' R&D investments.

TABLE I

*Real Growth in Energy R&D Investments, 1991–2001
(Percentages)*

Country	Change in energy R&D investment, 1991–2001
United States	–13
Japan	+15
France	+17.8
United Kingdom	–71
Italy	–50
Germany	–46.5
Canada	–43
The Netherlands	–14
Switzerland	–11

Note. The majority of major performing countries made large cuts in energy R&D support during the 1990s. As the table shows, restructuring and deregulation of the energy industries have been strongly correlated with energy R&D funding cuts. France and Japan are the only major energy R&D-performing countries that have not yet restructured their energy industries.

as holding significant promise as a future alternative to fossil fuel-based energy systems, investment in these areas has risen sharply during recent years. In Germany, fuel cell research has grown by 400% since 1998. In the United States and Canada, hydrogen programs have nearly doubled over the same period. The European Union and Japan also sponsor major

public and private sector hydrogen fuel cell programs. Similarly, government-sponsored biomass/biofuels, energy efficiency, and carbon sequestration programs have grown moderately in the United States, Japan, Germany, and Australia. Yet even with significant growth in these selected technology areas, emerging energy technology programs still constitute just a small share of energy R&D budgets.

Assessing private sector energy R&D trends has always presented difficult challenges because most energy R&D-performing firms consider these investment data sensitive and proprietary. The private sector data that are available are usually gathered via a brief survey instrument by industry associations or government agencies and are published only at an industry aggregate level. Trend data from these sources reveal that private energy R&D investments have declined even more precipitously over the past decade or so than have public sector energy R&D investments in several countries, including the United States (–63%), Germany (–70%), the United Kingdom (–90%), and Canada (–40%).

The energy R&D sponsored by the private sector is largely short term focused, aiming to develop new products or improve existing ones so as to bring them to the marketplace quickly. Long-term energy R&D, which entails larger financial risks and uncertain time horizons, has generally been supported by government agencies. In the United States, government currently sponsors more than 60% of

long-term energy R&D, whereas industry supports more than 80% of short-term research.

5. ASSESSING R&D TRENDS: METHODOLOGICAL CHALLENGES

It is important to note that all of the preceding data are based on surveys that use the term “energy R&D” in a conventional and narrow sense. The term is used to describe R&D activities that were explicitly carried out by governments or industry primarily to increase or create new energy supplies or to use energy more efficiently and cleanly. This means that these data sets will exclude R&D whose main and *a priori* purpose was to create something like a new military system, even though at times strictly defined “military R&D” has spilled over to benefit the energy industry. The most obvious example of this spillover phenomenon is the aforementioned combined cycle gas turbines, which have had a profound impact on the energy industry and are largely derivative of military jet engines.

The exclusion of these ancillary R&D efforts is driven by practicality and a desire to keep some temporal integrity to the data set being examined here. That is, only in an *ex post facto* sense can one identify these technologies that spill over into “energy R&D,” and the selection of these technologies is an inherently subjective exercise. It is more feasible to track R&D expenditures that were initiated due to their relevance to energy. Analysts and decision makers who rely on energy R&D statistics need to understand this important nuance in the data on which they are relying.

Another important methodological issue that analysts must consider in assessing energy R&D data across countries concerns the exchange rate mechanism used to convert the data to a common (typically U.S. dollars) currency. Many publicly available, multicountry energy R&D data sets use market exchange rates rather than purchasing power parities for their conversion to U.S. dollars. This method can introduce significant distortions into the data, often inflating actual energy R&D investments. Given that energy R&D is nearly exclusively funded and performed within—rather than across—national borders, exchange rates based on purchasing power parity yield more accurate estimates than do market exchange rates, which are more appropriately used for internationally traded goods and services.

It is also important to note that the use of brief survey instruments, by the IEA and others, as a primary mechanism for energy R&D data gathering should instill a measure of caution in analysts working with data sets compiled in this manner. Such survey instruments typically contain questions regarding a country’s or firm’s overall spending levels on energy R&D in selected program areas but offer little detail below aggregate levels. The use of survey instruments in this manner also raises questions regarding data accuracy and reliability given that survey participants are usually given few if any guidelines regarding their responses. Consequently, investments in fossil energy efficiency could, for example, be classified as “fossil energy R&D” by one participant and as “energy efficiency R&D” by another. Finally, verification of energy R&D statistics provided to international agencies via surveys is not possible. Thus, although these statistical data sets may be some of the best available, they must nonetheless be used and interpreted with caution.

6. CONCLUSION

Energy R&D is an important vehicle for the achievement of a wide variety of societal and policy goals. Because energy technology and fuel choices are critical determinants of economic prosperity, environmental quality, and national security, energy R&D is an important public policy vehicle. Moreover, the relative strength of energy R&D support is also a useful metric for the assessment of a number of public goods. The extent to which firms and government agencies view energy R&D as a priority activity and fund it accordingly may be regarded as a proxy indicator of political interest in selected policy areas such as climate change mitigation, national energy security, and overall environmental stewardship.

In this regard, it is important to note that available energy R&D data sets provide an important, but limited, measurement of inputs into energy R&D and offer little insight into R&D outputs—new technologies, their deployment, and their impacts on the environment as well as the economy and other policy concerns. Other common proxy measures of R&D efforts, including statistics on patent grants and applications, doctoral degrees awarded, and trends in energy intensity, can serve as useful pieces of the broad mosaic of energy R&D data but can be difficult to interpret outside a broader analytical context. Like statistics on energy R&D expenditures, these data are used as proxy measures because energy

R&D statistics are relatively easy to collect and have the appearance of being easy to interpret. In fact, output measures such as energy intensity (i.e., energy input per unit of GDP) are notoriously difficult to interpret because the number of variables influencing observed trends is potentially large and the isolation of individual variables is often impossible. For example, given data limitations, it might not be possible to determine definitively whether improvements in industrial energy intensity are due to the introduction of more fuel-efficient energy technologies, to broader changes in production methods, to structural changes in the economy, or to any number of other factors.

Most recent studies of trends in energy R&D have documented significant reductions in the overall level of support and a concomitant redistribution of remaining energy R&D funds in favor of less risky, short-term evolutionary R&D with clear commercial applications rather than longer term, higher risk projects with potentially revolutionary impacts. Many analysts have interpreted this trend as inherently dangerous given the likelihood that such funding patterns are likely to delay or prevent the attainment of many energy policy goals such as climate change mitigation. On the other hand, some analysts have pointed out that a reduction in the amount spent on energy R&D in itself cannot always be interpreted as a step backward for a firm or for society at large. For example, it is possible that energy R&D-performing laboratories might have become more efficient in their delivery of energy R&D results due to the increasing use of computer modeling or due to mergers between firms with large and redundant R&D capabilities that are subsequently consolidated—a frequent occurrence in the oil and gas industries. In either case, energy R&D statistics would reflect the fact that smaller amounts are invested in energy R&D, yet the implications of these events for energy technology innovation would remain indeterminate.

A final caution concerns forward-looking statements regarding future patterns of energy R&D investment and performance. As all investors are warned, past performance cannot be relied on as a predictor of the future. Many factors, including economic downturns, unforeseen energy crises, or environmental surprises can and have affected energy R&D investment in significant and unpredictable ways. Likewise, policy rhetoric regarding energy R&D priorities often remains disjointed from actual

R&D expenditures. For example, during the 1990s, many governments issued numerous white papers, policy briefs, and press releases about the importance of reorganizing their energy R&D portfolios to focus on climate change or to put more emphasis on renewable energy R&D. However, the historical budget data that detail how these governments actually allocated their energy R&D funds reveal modest changes in overall energy R&D emphasis. Thus, for the energy R&D policy analyst, *caveat emptor* remains sound advice. Analysts are advised to seek out multiple sources of qualitative and quantitative data on energy policy and energy R&D before drawing conclusions about the health and direction of the energy R&D enterprise.

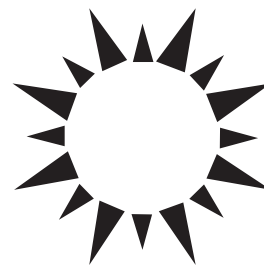
SEE ALSO THE FOLLOWING ARTICLES

European Union Energy Policy • Experience Curves for Energy Technologies • Innovation and Energy Prices • Microtechnology, Energy Applications of • Technology Innovation and Energy

Further Reading

- Dooley, J. J. (1998). Unintended consequences: Energy R&D in a deregulated energy market. *Energy Policy* 26, 547–555.
- Dooley, J. J., and Runci, P. J. (2000). Developing nations, energy R&D, and the provision of a planetary public good: A long-term strategy for addressing climate change. *J. Environ. Dev.* 9, 215–239.
- Electric Power Research Institute. (1996). “The Impact of Electricity Restructuring on Research for the Public Good,” BR-107160. EPRI, Palo Alto, CA.
- Lichtenberg, F. (1987). Changing market opportunities and the structure of R&D investment: The case of energy. *Energy Econ.* 9, 154–158.
- Mansfield, E., and Switzer, L. (1984). Effects of federal support on company-financed R and D: The case of energy. *Mgmt. Sci.* 20, 562–571.
- Margolis, R. M., and Kamen, D. M. (1999). Underinvestment: The energy technology and R&D policy challenge. *Science* 285, 690–692.
- Pacific Northwest National Laboratory. (2000). “Energy Research and Development: Global Trends in Policy and Investment.” <http://energytrends.pnl.gov>.
- President’s Council of Advisors on Science and Technology. (1996). “Report to the President on Federal Energy Research and Development for the Challenges of the Twenty-First Century.” PCAST, Washington, DC.
- President’s Council of Advisors on Science and Technology. (1999). “Powerful Partnerships: The Federal Role in International Cooperation on Energy Innovation.” PCAST, Washington, DC.

Resource Curse and Investment in Energy Industries



PAUL STEVENS

University of Dundee
Dundee, United Kingdom

-
1. A History of Interest in Resource Impact
 2. The Presence of Resource Curse
 3. Transmission Mechanism between Large Resource Revenues and Poor Performance
 4. How Resource Curse May Be Avoided: The Theory and the Practice
 5. Conclusion

Glossary

crowding out Where one sector manages to secure first call on resources, preventing access to those resources for other sectors.

developmental state A country in which the government espouses an ideology of development and tries to create the capacity to implement policies to promote development.

Dutch disease As a result of large oil or gas revenues, the real exchange rate appreciates, causing imports to rise and non-oil or gas exports to fall. The result is a contraction in the non-oil or gas tradable sector.

oil fund An institution whose function is to receive the revenues from oil or gas projects and to control their use in terms of both investing and spending.

predatory state The opposite of a developmental state; the government is a kleptocracy in which the ruling elite is out to plunder the wealth of the country for its own selfish purposes.

rent seeking Where managers in an enterprise—state or privately owned—seek to absorb resources to improve their own working environment and lifestyle. This is not the same as corruption since rent seeking is viewed as normal and legal behavior.

terms of trade The amount of goods or services that have to be exchanged for other goods or services; for example, the number of agricultural products that must be exchanged for an amount of manufactured goods.

Large revenues from natural resources should generate wealth, promote economic progress, and reduce poverty. The logic behind this may be based on common sense (more money may be better than less) or economic development theories, including capital fundamentalism, dual-gap analysis, or the big push argument to break out of a self-feeding circle of poverty. Thus, large windfalls from oil or gas projects should overcome capital and foreign exchange constraints and generate the required “push.” However, for many years casual empiricism and more formal investigation suggest the opposite. Many resource-rich countries have experienced a worse performance in economic progress and poverty reduction than countries without such apparent benefits. In the economics literature, this has become known as the resource curse thesis. However, it should be noted that although many suffered a “curse” from the influx of large natural resource revenues, others did not. Thus, resource curse is by no means the default situation.

1. A HISTORY OF INTEREST IN RESOURCE IMPACT

Concern regarding the impact of natural resources dates back many centuries, most obviously with Spain’s wealth from the New World. Concern emerged among development economists such as Raul Prebisch and Hans Singer in the 1950s and 1960s. They argued that primary product exporters (the periphery) would find themselves disadvantaged in trading with the industrialized countries (the center) because of deteriorating terms of trade.

Concern was focused on primary products and natural resources generally. However, in the 1970s, attention focused on the oil exporters following the oil

shocks. Their aftermath caused speculation that large-scale revenues might be bad for their development prospects. The experience of The Netherlands following the discovery of the Groningen gas field gave rise to concern regarding Dutch disease. This was defined explicitly as the contraction in the nonhydrocarbon-traded sector following a real appreciation of the exchange rate. For The Netherlands, this meant a decline in manufacturing. However, economists also began to observe that for oil-exporting developing countries, it was agriculture that took the hit.

In the 1990s, it was the impact on government behavior that attracted attention. Large windfall revenues from oil or gas projects appeared to change government behavior, damaging growth and development. This ran parallel to the wider developments in economic theory on the economic theory of politics, theories of public choice, and principal-agent analysis that laid the ideological basis that government involvement in the allocation of resources would lead to misallocation.

Recently, several factors revived interest in the subject and also brought resource curse to a wider audience. For some time, nongovernmental organizations (NGOs) have expressed concern that oil or gas projects in poor developing countries have negative effects. The World Bank Group set up an "Extractive Industry Review" to consider the bank's role in such projects. The larger, more responsible oil and mining companies are also considering their impact on host countries and how to mitigate negative consequences. This is driven by concern for reputation in a world in which ethical considerations of investment are becoming paramount. It is also driven by a view that negative effects could threaten the economics of the project.

Another cause of this revival of interest is that a number of countries—Azerbaijan, Angola, Chad, Kazakhstan, Papua New Guinea, and numerous others—will soon receive large-scale oil or gas revenues. There is general concern about the impact and use of these revenues and what should be done to maximize the benefits and minimize negative effects.

Finally, many institutions financing such extractive projects realize that the potential for resource curse from the projects threatens the viability of the projects and, hence, ultimately the value of their investments.

2. THE PRESENCE OF RESOURCE CURSE

A large body of work has attempted to establish a negative relationship between the abundance of

natural resources and poor gross domestic product (GDP) performance. For the most part, the evidence appears to support the negative link. However, such work is sensitive to the period chosen and there is evidence that before the 1970s resource-abundant countries grew faster than resource-deficient ones. Also, because of price fluctuations certain periods are likely to distort results.

There are a few dissenting voices to this empirical work based on methodology rather than the overall conclusions, although some challenge both methodology and conclusions, arguing that there is little long-term evidence that natural resource-abundant countries generally underperform. Rather, natural resources have played an integral role in the success of many industrialized countries. Oil exporters have also provided support for resource curse. For example, Mikesell identified Saudi Arabia, Mexico, and Venezuela as having lower than average annual growth rates of GDP.

Apart from a negative growth impact, another aspect is the prevalence of poverty and the observation that oil- or gas-rich economies have a poor record in poverty alleviation. This is a central theme of much of the literature. Another dimension is that oil or gas projects are associated with greater conflict in a society. Large-scale resource revenues create a pot that is worth fighting for. Whoever is in power is better able to plunder that pot. Such revenues also generate higher levels of military spending. The projects often alienate local populations, especially if separatist tendencies already exist. Conflict is also linked to poverty. War and strife (e.g., inflation) are regressive since they hurt the poor more than the rich because the poor cannot mitigate against the effects. Fighting or expectations of fighting also absorb resources that could go to improving economic performance and alleviating poverty.

Another characteristic linked to resource curse is that natural resource abundance slows political change and entrenches regimes. Thus, as Auty noted, an abundance of resources significantly weakened nascent democratic institutions, repressing political parties so that power is weakly contested, public finances are opaque, and corruption by both the elite and the bureaucracy is rampant.

A final dimension is the regional impact of the projects. Although the effect at a national level might be debated, because of the local impact, damage is often done especially in terms of both the environment and human rights. Meanwhile, the benefits appear to flow to central rather than regional authority.

However, although evidence of the presence of resource curse appears convincing, there are examples of countries for which there appears to have been a “blessing” rather than a curse, including Botswana, Chile, Malaysia, Norway, Indonesia before 1997, Australia, and Canada. A study by Stevens provides strong support for the fact that these countries avoided the curse. Thus, there is a case for dropping the automatic use of the term resource curse and instead using the term resource impact and considering whether the outcome was a curse or a blessing.

3. TRANSMISSION MECHANISM BETWEEN LARGE RESOURCE REVENUES AND POOR PERFORMANCE

There is a huge literature on why countries might suffer a curse rather than a blessing following a large inflow of oil or gas revenues.

3.1 Long-Term Decline in Terms of Trade

This argues that over time, the revenue would buy less imports of capital goods, thereby inhibiting development-creating investment. However, this is controversial and the empirical and theoretical arguments have been much contested. Also, it is difficult to understand how a slow decline in prices would explain the sort of deterioration in economic performance associated with resource curse. Economies constantly face changing relative prices. If they take place gradually, economies might be expected to adjust unless other factors are at work.

3.2 Revenue Volatility

Oil or gas revenues are very volatile, especially driven by wide fluctuations in prices over relatively short periods of time. Certainly, there is plenty of empirical support for the existence of such volatility. Thus, Mikesell found that between 1972 and 1992, regions with high primary export shares experienced periods of trade volatility two or three times greater than those of industrial countries during the same period. Fluctuating revenue profiles make it difficult to pursue a prudent fiscal policy. This creates problems ranging from aggravating investor uncertainty to “stop go” spending policies. There is also concern that windfall revenues would be consumed rather

than invested, although empirical evidence throws doubt on this since the permanent income hypothesis argues that windfall gains would be more likely to be saved and invested than consumed. However, intuitively it is easy to imagine fluctuating revenues, in the absence of effective stabilizing measures, creating problems for government fiscal policy and macroeconomic management more generally.

3.3 Dutch Disease

Originally, this referred to the appreciation of the real exchange rate as a result of inflation from spending the revenues plus an appreciation of the nominal exchange rate as the domestic currency attracted higher demand. The result was a contraction in the non-oil or gas sector. However, over time, the meaning has become wider to encompass all negative macroeconomic effects associated with resource curse.

These changes of meaning are reflected in the differing theoretical approaches. Corden and Neary split the impact of an oil boom into a resource movement effect and a spending effect. In the former, a higher marginal product in the booming resource sector draws resources from other sectors. Thus, factors move into oil, bidding up wages and causing other sectors to contract. The spending effect occurs when, as a result of the windfall, demand rises in both tradables and nontradables. Since prices of tradables are determined by the international market, greater demand is met by higher imports. However, prices of nontradables increase relative to tradables and so resources shift from tradables to nontradables. This approach forms the basis for much of the subsequent analysis in the literature.

Over time, other dimensions of Dutch disease—defined simply as a contraction of the tradable sector—have emerged. Thus, one argument is that subsidies used to protect nonresource tradables that are weakened by the boom aggravate the sector’s problems and eventually become unsustainable. Another is the “leapfrog effect,” in which governments tend to miss the labor-intensive phase of industrialization and move straight to the heavy, capital-intensive phase with negative effects for the tradable sector. It has also been suggested that resource-poor countries accumulate social capital faster than resource-rich countries. One explanation for this is that limited natural resources promote early industrialization forcing earlier urbanization.

In addition to the controversy regarding the exact nature of the theory of Dutch disease, there is also

controversy regarding whether empirical evidence supports its existence. A number of country studies provide support, including studies on Bolivia, Venezuela, Mexico, Brazil, Peru, Algeria, Ecuador, Indonesia, Nigeria, and the United Kingdom. However, often the situation is complicated by extraneous factors. For example, in Botswana a Dutch disease effect on agriculture was complicated by two serious droughts in the 1980s. Others conclude that Dutch disease was not the major factor explaining the growth pattern and that some deindustrialization may be the result of adjustment toward a new equilibrium. Furthermore, for oil-exporting countries, after the oil boom of the 1970s, although agriculture did contract, manufacturing tradables actually expanded in most cases.

3.4 Crowding Out Effects

This is a variation of the resource movement effect of Dutch disease in which the project, which is large relative to the rest of the economy, stakes first claim on scarce resources. The rest of the economy finds it difficult to secure the factors needed to develop. This is especially relevant in smaller countries when the size of the investment project is large or where there are strong regional dimensions to an economy and the project is focused in one region. It is also important in the transition economies, in which transition has destroyed the old productive base and a new and private tradable sector is struggling to emerge.

3.5 Increasing the Role of the State

If the “Washington consensus” is believed, then greater government intervention in an economy is a bad thing. Given that in most legal regimes the revenues accrue to the government, this inevitably invites greater government intervention. Many of those writing on resource curse view a major part of the explanation of the phenomenon as relating to the role of the government. Also, many argue that it was good government decisions that enabled a few countries to avoid the worst excesses of resource curse. As with crowding out, the issue is especially sensitive in the transition countries simply because the objective of transition is to reduce government intervention in the economy.

However, although many would agree with these sentiments, why oil- or gas-exporting countries appear more vulnerable to policy failure requires further explanation. Several strands of argument are possible. The first argues that large windfalls lead to

poor general decision-making by governments. The development of such projects raises expectations. This pressures government to “do something” that encourages speedy responses. Often, quick, ill-coordinated decisions are bad decisions. Also, spending revenues too quickly is more likely to introduce distortions into the economy, if only because there is less chance for the economy to adjust naturally. Having more money to play with weakens prudence and normal procedures of due diligence. Thus, making the “right choices” seems somehow less important. Of particular relevance is when governments decide on capital spending without consideration of the recurrent spending implications. Finally, because in the first instance the revenues accrue to government, decision-making is concentrated in fewer hands than if a greater number of economic agents are involved in decisions regarding how to spend any windfalls.

Another strand of explanation concerns corruption and rent seeking. Having large amounts of money to play with increases the temptation for corruption and rent seeking on the part of the decision makers. However, this is a complex and controversial area. Some argue that corruption evolves from a clash between traditional values and foreign norms. Thus, in Africa, bureaucrats are under an obligation to share the benefits of their office with their kin, a characteristic common to a great many other societies. If this is the case, then because the oil or gas projects in developing countries almost invariably involve foreign companies, this alone helps explain why resource-rich countries may experience greater levels of corruption.

Rent seeking concerns how people compete for “artificially contrived transfers.” The result is that the expenditure creates no social value and, arguably more important, distorts markets and hence the way the economy operates. There are several explanations for why rent seeking is greater in countries with large oil or gas revenues. Most obvious, the larger the public purse, the less noticeable the leakage to interest groups, although this may not be true if the result is a “feeding frenzy.” Also, rent seeking is greater because wealth is concentrated in the public sector or possibly a small number of companies. Thus, most of the rents created are channeled by bureaucrats, most of whom are members of the politically dominant group. There is general agreement that rent-seeking behavior produces undesirable results for the economy and imposes significant losses on many economies. Such behavior distracts attention from goals of long-term development

toward maximizing rent creation and capture. Also, rent seeking creates powerful lobby groups that block much needed economic reforms. Rent seeking also makes it more difficult for governments to adjust spending when faced with revenue fluctuations.

However, such issues are far from clear-cut. An important aspect is what happens to the rewards of corruption and rent seeking once they have been secured. If used for productive investment within the economy, the outcome will be different than if used for conspicuous consumption or to enhance foreign bank accounts. After all, at the risk of being simplistic, the whole point of capitalism is for a few greedy, selfish, and unscrupulous individuals to “grab” resources and put them to productive use to secure even more resources. The “robber barons” who developed the U.S. economy in the 19th century (very resource rich by any definition) indulged in some very dubious business practices, not least in the oil sector.

Often, the nature of investment decisions following large windfall revenues creates problems for economic progress. This is in part explained by the bad decision-making previously discussed. Also, often investments made by resource-rich governments fail to develop the productive base of the economy. Furthermore, if borrowing funds these investments, this compounds the problem since the loan must be serviced and repaid. It has been argued that much of the link between resource abundance and negative growth may be attributed to the aftereffects of a borrowing spree by natural resource exporters in the 1970s.

A specific area that combines poor decision-making and bad investment decisions in many resource-rich countries concerns the industrial policy adopted. Resource-rich countries tend not to promote a competitive manufacturing sector. Rather, they adopt an industrial policy based on import substitution involving subsidy and protectionism. Often, resource revenues prompt the introduction of heavy subsidies, usually based on an infant industry argument. These become unsustainable when revenues decline. However, their creation and the subsequent rent seeking create powerful interest groups that oppose their removal. Regarding protectionism, there is a positive correlation between dependency on primary products and a closed trade regime. Once these elements of policy—subsidy and protection—are in place, continuing resource revenues reduce the incentive to create competitive manufacturing industries. Since many development economists regard competitive manufacturing as a principal source of

technological progress, this has serious implications for economic progress.

3.6 Sociocultural and Political Impacts

The sociopolitical dimension is central to the discussion, although its role is much disputed. It has been suggested that resource-rich countries have weaker institutional capacities, and that the endogenously determined institutional capacities are key determinants of growth. In contrast, others argue that although differences in income can be explained by the effectiveness of institutions, this is a legacy of colonialism and there is no discernable impact from natural resource abundance.

Previous explanations related to the macroeconomic causes of the curse. Countries that avoided the curse did so by the adoption of specific policies to minimize damage. Of key interest is why such policies were or were not adopted and why the various political elites allowed them. The answer lies in the nature of the society and its politics and, as such, is a crucial part of the explanation of the transmission mechanism.

Some of the discussion revolves around rentier societies in which there is no connection between reward and effort. Thus, easy access to riches bred idleness and provided no incentive to accumulate or innovate. Although this explanation seems attractive, it is not conducive to analytical rigor and smacks slightly of stereotyping that, apart from any other potential problems, condemns the recipients to a life with no hope of redemption.

A particularly fruitful area of sociopolitical enquiry concerns the categorization of states as developmental or predatory. A developmental state has two elements—an ideological and a structural component. The ideological component involves the ruling elite adopting developmentalism as its prime objective. Legitimacy is derived from the ability to deliver development, implying both growth and poverty reduction. This elite establishes a form of ideological hegemony—via the ballot box or less desirable means—over the rest of the society. The structural component involves the capacity to implement policies “sagaciously and effectively” that will deliver development. Apart from technical capabilities, this also requires a strong ruling elite that can resist pressure from powerful, short-sighted private interests and/or some form of “social anchor” that restrains any temptation to use its autonomy in a predatory manner. A key element is the realization that developmental states can still fail. Despite the

“right” ideology and the limits of predation in place, the capacity to implement effective policies might not be enough to manage certain problems, especially if these are exogenous. Such problems may be driven by external shocks, mistakes, or bad luck.

In contrast, in the predatory state, the ruling elite only have interest in feathering their own nests. They are “takers,” not “makers,” who have little or no interest in promoting economic development. Thus, their objective is to plunder the state’s resources to benefit the few, as opposed to a developmental state in which the objective is to benefit the many. There are no checks and balances to inhibit plundering the national wealth. Indeed, the less constraints on the ruling elite, the more they can plunder. This process becomes self-reinforcing. The greater the wealth accumulated, the more resources they have to inhibit, buy off, or control any potential checks and balances.

Large-scale oil revenues encourage predatory states by a number of mechanisms. As with rent seeking and corruption, they make the pot larger and therefore more attractive to take. The more there is to steal, the greater the temptation to do so. Revenues generate resources to suppress opposition to the predation either by the use of force or by simply buying off potential problems. Finally, the major powers, dependent on oil, may provide support and legitimacy for predatory regimes to ensure the continued flow of oil, thereby protecting the regime both from internal problems and from threats from neighbors. Iraq is a classic example.

It is tempting to attribute poor performance by resource-rich countries to a lack of democracy. However, the argument is debatable. Thus, as Mikesell suggested, democracy does not ensure good government, nor are all oligarchies poorly governed. Furthermore, the democratic credentials of countries that avoided the curse—Malaysia, Indonesia, Pinochet’s Chile, and indeed Botswana—are highly debatable.

This lack of clarity on the transmission mechanisms presents a serious challenge. If it is not clear what causes the problem, then treatment is likely to prove equally elusive. However, this is important if investment in oil and gas projects under consideration is to go forward.

4. HOW RESOURCE CURSE MAY BE AVOIDED: THE THEORY AND THE PRACTICE

The literature presents a number of options for avoiding resource curse.

4.1 Leave It in the Ground

Although this cannot be regarded as a serious option since a less ethical company will always be willing to make money, irrespective of damage or reputation, it does raise issues about the speed of the project and hence the consequent production/revenue profile.

Conventional project economics argues that from a commercial standpoint, the faster the development, the better. The slower the production, the lower the total present value of the revenue stream. However, common sense argues that a slow, steady trickle of revenues should be easier to handle than a sudden surge. If revenue management problems are easier to deal with, the resource movement effect and crowding out are likely to be less severe. Furthermore, slow development allows the development of a service industry to maximize backward linkages from the project, whereas swift development must be based on imports. The problem is persuading investors to slow development. Although this goes against conventional project economics, if rapid development generates a curse and if this threatens the project because of the political fallout, slower development may well be justified on purely commercial grounds. There is a major gap in the literature on resource curse regarding how the phenomenon might affect the economics of a project. One obvious connection is that insofar as the curse embodies increased conflict within a country, this may directly physically threaten the project’s equipment and infrastructure. This would be especially relevant where the curse had a strong regional dimension as a result of the local impact of the project.

A variant on this theme is the nature of the revenue flow. Differing forms of extractive agreements—concession based on royalty and taxes, joint ventures, production-sharing agreements, service agreements, etc.—will create differing profiles of revenue flow. Thus, although development of the project may still be rapid, the revenue flow could be controlled by the agreement rather than leaving it to the vagaries of some sort of stabilization fund. An example is the case in which a project is expected to create a short-lived revenue peak for the government. In this case, the project may flatten the peak by providing revenue upfront before the project is fully operational and afterwards when the project is finished. Of course, this implies that it is the company that controls the revenue flow rather than the government, which raises issues of sovereignty.

4.2 Diversification

One obvious solution is to reduce the importance of the oil or gas sector in the economy by diversifying the economy. Since the early 1970s, oil-exporting countries have paid lip service to diversification away from dependence on crude oil exports. However, the record in general has been very poor, with huge amounts of public money being poured into inefficient and uncompetitive industries.

Several factors explain these failures. Dutch disease and crowding out cause a contraction of an existing non-oil or gas tradable sector. It is not surprising if the mechanisms that destroy do not also inhibit the rise of a new tradable sector. Another explanation favored by countries is that trade restrictions imposed by the Organization for Economic Cooperation and Development exclude their outputs. Although both arguments have validity, there is a third explanation. In most cases, the diversification policy has consisted of governments trying to pick winners. Thus, much of the new economic activity has been government-owned enterprises. Governments are generally very bad at picking winners, although there have been exceptions.

The only effective diversification derives from private sector investment, although governments can play an important role in this process, as can be seen from the experience of the Asian tigers. Thus, one solution to the problem of diversification is to maximize the revenue flow to the private rather than the public sector (discussed later). This was certainly an important part of the explanation for why Australia, Canada, and the United States avoided the curse. Of course, this raises important issues of income distribution and private economic power leading to concentrations of political power. This in turn is associated with the complex and controversial issue of how much political repression is worth accepting for an improved standard of living—a constant theme of more general literature for centuries.

Another means to encourage successful diversification relates to trade policy. Countries that avoided the curse pursued a policy of trade openness with the rest of the world coupled with a deliberate policy of exchange rate depreciation to ensure the competitiveness of non-oil or gas exports.

4.3 Revenue Sterilization

A key factor is the macroeconomic policy adopted by the government, specifically the policy adopted to try

to neutralize the impact of the large windfall revenue inflow on the rest of the economy. This has several dimensions. The first is the need for fiscal prudence to prevent the revenues from translating immediately into greater aggregate demand and inflation. This requires the government to resist spending pressures and either accumulate budget surpluses or channel the revenues into some form of fund. An important subissue is that the government should avoid using promises of more revenue to increase borrowing.

Another important strand to macropolicy is to avoid exchange rate appreciation. Several options are available. The first is to avoid a fixed exchange rate policy when domestic inflation of any magnitude exists. The second is to encourage currency devaluation. However, this has no effect on the real exchange rate without appropriate demand management policies. One option is for the central bank to purchase foreign exchange to hold down the nominal rate. One common factor in the case of countries that recently avoided a curse is that they all experienced significant depreciation of the real exchange rate as a result of explicit policy choices.

4.4 Stabilization and Oil Funds

There is general agreement that trying to stabilize spending to ensure stable and moderate growth is important. This can be done in several ways. Trying to stabilize the international price is no longer a serious option and probably never was despite a large literature on commodity stabilization schemes. One possible solution is to stabilize prices by using futures markets to lock in price stability. However, there are problems. For very large producers, the danger is that their involvement would simply swamp the markets. Also, such insurance costs. Although having guaranteed prices in a falling market would result in praise for the decision makers, in a rising market there is bound to be criticism. Another solution might be to use the extractive agreement to smooth revenue flows, as previously suggested.

Another mechanism is to create an oil fund. These can fulfill three functions. They can be used to insulate the economy from large revenue windfalls by investing them outside the domestic economy. The fund can also be used to stabilize revenues by setting a price assumption for budgetary purposes. If prices exceed this, then the fund absorbs the windfall. If prices are lower, then the fund's assets are used to meet the budget. Finally, such funds can be used to put assets aside for future generations. Thus, the

revenue is accumulated and only the earned income is spent, leaving the capital sum intact to allow future generations to benefit from a depletable resource.

In recent years, the role of such funds has become extremely controversial, especially within and between the World Bank and the International Monetary Fund. There are two schools of thought. One argues that the conditions for such funds to work are such that if they are present, the revenue windfall can be managed within the normal budgetary process. Hence, the funds are not needed. The other argues that even if conditions are not ideal, the creation of such funds makes some contribution to avoiding the worst excesses of resource curse if only because of the role of the fund in dampening spending expectations.

4.5 Investment Policy

The investment policy adopted by the government for the revenues will play a crucial role both in avoiding many of the macroeconomic pitfalls described previously and in encouraging the process of economic diversification by creating alternative sources of value added to the depleting resource sector. Ideally, investment is best carried out by the private sector. Although, as previously noted, the experience of the robber barons in the United States in the 19th century illustrates that this does not guarantee everyone lives happily ever after. However, private sector investment is more likely to be sustainable than when the government tries to pick winners.

4.6 Political Reforms Needed to Carry Out the Corrective Politics

A key component of any solution is the need to remove corruption and contain rent seeking. However, this is far from straightforward. For example, any cleanup campaign must start with an agreement on a common standard of morality. Options to solve problems of corruption include incorporating controls in legal codes and encouraging independent (and fearless) enquiries. Another solution is to reduce rent-seeking opportunities by depoliticizing resource allocation. Thus, deregulation allows free markets, taking politicians and bureaucrats out of the resource-allocation process. The theory with empirical support is that competition squeezes out illicit margins. Although this is probably true, it still requires the creation of competition, and often simple deregulation and privatization is not sufficient to achieve this.

Much is also made of the need for transparency as an aide to reduce corruption. Although this can have a role, there is a danger that its effectiveness can be overstated. For example, many have called for the oil companies to disclose how much money they provide to the government so the public can see how much money is being “diverted.” This neglects the fact that often very large sums accrue to the national oil company as a result of the agreements, especially joint ventures and production sharing agreements. It also bestows upon both the public and the local media a degree of sophistication that may be misplaced. For example, a major problem is that because the cash flow associated with projects is uneven over time, considering only a few years could give the impression that the oil companies are taking a huge share of the proceeds, ignoring the earlier years of negative cash flow. This could revive the “obsolescing bargain” and lead to unjustified pressures to renegotiate terms. This is not to argue that there is no place for transparency, just that it is far from a sufficient condition to reduce corruption.

There is also the issue discussed previously that from an economic standpoint, it is not so much the presence of bribes and rent that is the problem. Rather, it is how they are used. Leaving aside moral issues and issues of income distribution, bribes and rent productively invested result in a successful capitalist economy. Economic problems arise when these are diverted into conspicuous consumption or foreign bank accounts.

Following from previous discussion, a key issue relates to securing a developmental rather than a predatory state as a way of minimizing the risk of suffering a curse. Developmental states require two components: a strong elite with an ideology that is conducive to economic development and whose interests are aligned with the population and also the institutional capacity to implement the necessary policies without being captured by narrow private interests. In contrast, a predatory state lacks the linkages to the population or any other constraint that prevents the elite from ransacking the economy.

Securing an alignment of interests is crucial. One route to achieve this alignment is democracy in which the government must satisfy public interests or be voted out. However, although this may seem to be the optimal route morally, as argued previously, it is not the only route. Thus, for example, Chile under Pinochet and Malaysia, Indonesia, and indeed Botswana as a “one-party democracy” managed the alignment quite well.

Of course, what constitutes an ideology that is conducive to economic development is also a moot point. Clearly, simple mantras based on “leave it all to the market” are inappropriate. However, the experience of those countries that avoided a curse suggests that although their governments did intervene extensively, they did so to push the economy toward a market-based system linked into the global economy and toward the encouragement of a vibrant private sector.

The ability of the elite to translate this ideology into effective policies is a function of institutional capacity. Clearly, both international financial institutions (IFIs) and investing companies have a role to play in training those elements of the bureaucracy that could play a key role in policy making. This can be done using a combination of seconding expertise (assuming acceptance by the governments) and training nationals using short professional training courses and more long-term basic provision of university postgraduate training.

All this is feasible and can be made to work such that the economy progresses. The real problem arises when the ruling elite lacks any social anchor and alignment with the public and the state has become predatory. Insofar as the state may be dependent on IFI good will for financing and debt management, the IFIs can be in a position to provide the constraints required to reign in predatory behavior, although this obviously raises sensitive issues of sovereignty. However, such an outcome has the seeds of its own destruction since once the resource revenues begin to flow, the IFI leverage seeps away and with it the constraint. One possibility is that the strengthening of civil society might begin to provide both the constraints and the alignment necessary to switch the state from being predatory to developmental. Again, the IFIs and companies can help strengthen civil society, but this means more than just giving the NGOs more money, although what precisely it does mean is uncertain and needs further thought and analysis. Also, it assumes the predatory state will tolerate the growing influence of civil society, which for obvious reasons of its own survival it may not. Another option is for the oil company to exert pressure on the government to curb predation, but this would only be effective if all the main oil companies were willing to take the same stance.

Of course, one solution is for the international community to refuse to invest in predatory states.

However, this raises the prospect of an unethical portion of the industry investing to add further to the plunder of an economy by its own elite. It also raises the prospects of creating ever more failed states harboring resentment against the West.

5. CONCLUSION

Clearly, the issue of what causes a curse rather than a blessing and how to enjoy the latter and avoid the former is extremely complex. The literature is divided on both causes and cures, although there is a growing consensus that essentially it has something to do with governance and the answer lies more in political economy than macroeconomic analysis. Hence, the key question is not what was done in terms of policy decisions but, rather, why was it done? The outcome of the debate will be of crucial importance in terms of determining the future magnitude, type, and direction of investment in the oil and gas sector.

SEE ALSO THE FOLLOWING ARTICLES

Development and Energy, Overview • Experience Curves for Energy Technologies • Investment in Fossil Fuels Industries • Oil-Led Development: Social, Political, and Economic Consequences • Oil Price Volatility • OPEC, History of

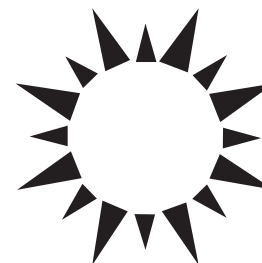
Further Reading

- Auty, R. (ed.). (2002). “Resource Abundance and Economic Development.” Oxford Univ. Press, Oxford, UK.
- Corden, W. M., and Neary, J. P. (1982). Booming sector and Dutch disease economics: A survey. *Econ. J.* 92.
- Meyer, J., Chambers, B., and Farooq, A. (eds.). (1999). “Development Policies in Natural Resource Economies.” Elgar, Cheltenham, UK.
- Mikesell, R. (1997). Explaining the resource curse, with special reference to mineral-exporting countries. *Resour. Policy* 23(4).
- Mkandawire, T. (2001). Thinking about developmental states in Africa. *Cambridge J. Econ.* 25.
- Sachs, J. D., and Warner, A. M. (1999). The big push, natural resource booms and growth. *J. Dev. Econ.* 59.
- Stevens, P. (2003). Resource impact: Curse or blessing? A literature survey. *J. Energy Literature* 9(1).
- Stevens, P. (2003). Resource impact: A curse or a blessing? *Electronic J.*, www.cepmpl.org.

Reuse and Energy

MARKO P. HEKKERT

Utrecht University
Utrecht, The Netherlands



-
1. Relationship between Energy and Materials
 2. Product Reuse in Perspective
 3. Examples of Product Reuse
 4. Environmental Benefits of Product Reuse
 5. Implementation Barriers of Product Reuse
 6. Conclusion

Glossary

improved material management Intentional actions that lead to a reduced amount of material that is used to fulfill the same function. Since materials require a large amount of energy in production, this often leads to energy savings.

material cascading Improved material management measure. It is a form of material recycling but differs due to the reduction in product quality that can be produced with the secondary material. An example of material cascade is paper used for printing and writing that is recycled for newsprint.

material recycling Improved material management measure; reuse of materials. The original product is destroyed in order to reuse the materials for new products. This can be the same type of product or a new type of product. Energy savings often occur since the manufacturing of virgin materials often requires more energy than producing recycled materials.

product reuse Improved material management measure. This is a type of material recycling, but it leaves the original product intact. The major advantage is that energy used in the construction of a product is not lost due to destruction of the product. It is the material management measure with the highest carbon reduction potential in the packaging sector.

Since the use of fossil fuels is an important cause of increased greenhouse gas concentrations, many strategies for reducing human greenhouse gas emissions focus on a reduced use of fossil fuels. To reduce the use of fossil fuels, a wide range of options can be

applied: (i) improvement of energy efficiency, leading to a reduction of the energy consumption per unit of activity or product; (ii) development and application of renewable energy sources, such as biomass, hydropower, solar, and wind energy; (iii) a shift in the use of fuels from resources with a high carbon content (coal) to resources with lower or even no carbon content (natural gas and uranium); (iv) decarbonization of fuels and flue gases, for example, by recovering carbon dioxide from energy conversion processes and storing it outside the atmosphere; and (v) improvement of material management.

This article focuses on reuse of materials, which can be categorized as an improvement of material management.

1. RELATIONSHIP BETWEEN ENERGY AND MATERIALS

Improvement of material management is not commonly incorporated in greenhouse gas reduction policies. This is strange since several studies indicate that this option has a large potential and is often economically attractive. Before the results of these studies are described in more detail, the reason why material efficiency may lead to reduced greenhouse gas emissions is discussed.

Materials are produced in the industrial sector and a major portion of fossil energy consumption is used in this sector. The industrial sector (excluding refineries) consumed approximately 33% of the total primary energy use in 1995 for developed countries. When materials are used more efficiently, less materials need to be produced and, therefore, less energy is needed in the production stage. Consequently, more efficient material management is likely to lead to reduced emissions of greenhouse gases.

A limited number of studies have been performed on the subject of improved material management and climate change. A few case studies from Utrecht

University on plastic packaging and fertilizers show that a significant greenhouse gas emission reduction is possible by more efficient material use. Patel studied a limited amount of material management options for the plastics sector in Germany. He showed that these options may lead to a 24% reduction of CO₂ emissions from the German plastics-producing industries. Furthermore, the U.S. Environmental Protection Agency showed that management of waste materials presents many opportunities for greenhouse gas emission reduction. Elliott showed that recycling, an improved material management option, may lead to major savings in industrial energy use. In The Netherlands, the MATTER project, involving the Dutch Energy Research Centre and several universities, was started to model the Western European energy and materials system and determine, based on technological improvement options, the greenhouse gas emission potential in these systems in the period 1990–2050. Gielen concluded, based on these model results, that “the potential for emission reduction in the materials system seems to be of a similar magnitude as the emission reduction potential in the energy system.” The importance of these studies for understanding the potentials for greenhouse gas emission reduction is stated by the United Nations Department of Policy Coordination and Sustainable Development. Based on these studies, one may conclude that the potential of improved material management on greenhouse gas emission reduction seems to be significant, and obtaining more insight into the exact potential and the ways to utilize this potential is important.

2. PRODUCT REUSE IN PERSPECTIVE

The environmental pressure that is related to a product or a material is the sum of the environmental pressures related to the production of that material or product as well as the other stages in the life cycle of the product, such as production of raw material and waste management.

Climate policy should therefore focus on greenhouse gas emissions at all stages of the product life cycle: raw material extraction, material production, product manufacturing, product use, and disposal of products and waste management.

Changes in any of these stages can lead to reduced product-related greenhouse gas emissions. However, it is important to realize that measures that lead to

reduced emissions in one part of the product life cycle may lead to increased emissions elsewhere: In order to reduce product-related greenhouse gas emissions, the whole life cycle needs to be taken into account.

Figure 1 gives an overview of the different life-cycle stages and the measures that can be taken to improve the efficiency of material production and use.

Figure 1 shows that in all stages of the life cycle, measures can be taken to increase material efficiency. The improvements can be categorized in many different ways. Here, the following classification is used:

1. Technical measures in the individual life-cycle stages
2. Management of the entire material product chain

Technical measures in the individual life-cycle stages include measures such as increasing the efficiency of processes (where innovations in technological development and process design can be used to reduce the amount of materials used in the process), improved product design [where using innovative production methods facilitates the use of thinner material (layers) and using other materials in the product manufacturing stage may lead to reduced emissions; important substitution criteria are the use

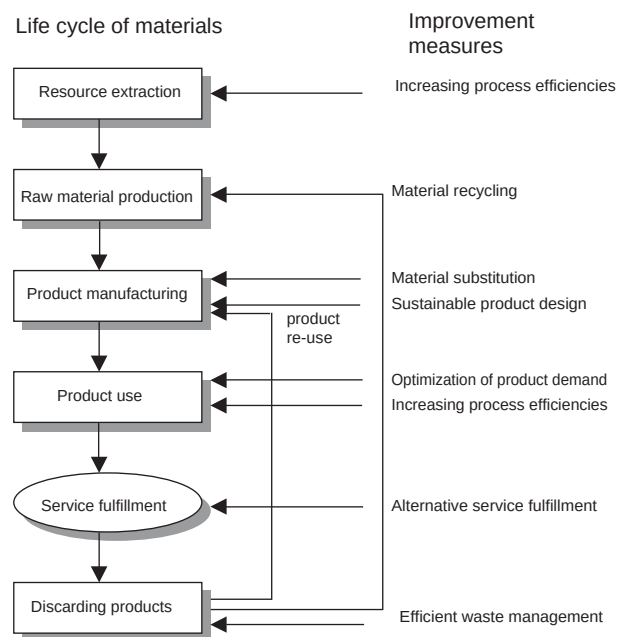


FIGURE 1 Material life cycle and improvement measures.

of energy-extensive materials, materials with a long lifetime, and materials that are suitable for reuse], and more efficient waste management (where there is optimal energy recovery from waste materials).

Management of the entire material product chain covers measures such as material recycling, material cascading (using discarded materials for the highest possible functions), extension of the useful life of products (modular design and increased fitness for repair and maintenance), sustainable product service systems (measures that lead to more efficient service fulfillment due to the substitution of products by services or combinations of products and services), and reuse (e.g., refillable plastic or glass bottles).

Reuse involves the collection and cleaning of products in such a way that they can be used again in the same form and shape as they were used in the previous life cycle. In such a way, both the structural properties of the material and the energy required for shaping the product are preserved. Therefore, to make a clear distinction between reuse and recycling, it is more appropriate to speak of product reuse and material recycling. In the case of material recycling, only the structural properties of materials are used again in new product life cycles. In this case, it is not necessary to use the recycled material for the original product. For example, polyethylene terephthalate (PET) bottles can be collected and recycled in two different ways. The first is to clean the PET bottle and reuse it as a bottle (product reuse). The second is to shred the bottle and recycle the PET material for a different product, such as crates or fleece jackets (material recycling). Material reuse and material recycling differ from material cascading by the fact that in a material cascade the product that is made out of recycled material is always lower in quality than that of the original product. Optimal material cascading policy tries to optimize the required material characteristics for a product and the factual material characteristics. An example of a material cascade is to produce a solid wooden beam for construction purposes from a tree, then recycle the material after its product lifetime as wooden floorboards or window frames, and recycle the material again after this life cycle as fiberboard. In this case, the required wood quality of the products declines over the consecutive life cycles. Another example is the use of virgin wood pulp for the production of high-quality printing and writing paper, reuse of the wood fibers for lower quality printing and writing paper, and finally use of the fibers for newsprint production.

3. EXAMPLES OF PRODUCT REUSE

To make product reuse possible, a good working collection system is necessary. Traditional waste collection systems are not sufficient since the original products are damaged and contaminated. Therefore, separate collection systems are necessary. There are a few examples in which these systems have been set up successfully.

The first example is the reuse of beverage bottles in several European countries. In The Netherlands, several types of beverage bottles are separately collected, cleaned, and reused. Examples are the reuse of 0.33-liter beer bottles (88% of the beer volume packed), 1- and 1.5-liter soda bottles (56% of the volume packed), and 1-liter milk bottles (10% of milk volume packed). The beer bottles are made out of glass, the soda bottles are made out of PET, and milk bottles are produced from polycarbonate (PC). In packaging terminology, every time a bottle is reused it is called a trip. Thus, the trip number is the number of times a bottle is reused. On average, these trip numbers are approximately 20 for beverage bottles, implying that 20 times less material is needed to fulfill the same packaging function.

The second example is the reuse of transport packaging, also popular in many European countries. Transport packaging is defined as packaging that is used to bundle and protect multiple products for transport purposes. Examples are corrugated transport boxes, crates, and pallets. The latter two are often reused.

Crates and boxes are normally used to pack loose products, such as fruit and vegetables, meat, and product parts. (Crates are also used for transportation of bottles. No improvements for these crates have been modeled.) The crates and boxes can be used for one-way shipping or function as a returnable (reuse) package. The one-way crate is mostly made out of wood or corrugated board and is normally used to pack fruit and vegetables. One-way boxes are always made out of corrugated board. This box has an average volume of 40 liters and weighs 0.8 kg. Multiple-trip plastic crates can be used instead of one-way crates and boxes. The plastic multiple-use crate is made from high-density polyethylene (HDPE), weighs 2 kg, and has a volume of 40 liters. Using plastic returnable crates requires a closed-loop transport system. A third party often manages the logistics of this transport system. The crates currently used have a lifetime of 5–10 years and the average trip rate is 25/year.

Pallets are also used as single-trip packaging and in reuse systems. A vast majority of the pallets used in

Europe are made out of wood. Wooden pallets are popular because they are cheap, have a large carrying capacity, and are easy to repair. Moreover, they are very suitable for production of small series with deviating sizes. Approximately one-third of the pallets used in Europe are returnable. A single-use pallet weighs approximately 17 kg and a multiple-use pallet weighs approximately 25 kg. On average, returnable wooden pallets make 20 trips. Obviously, much material saving occurs when returnable pallets are used instead of single-use pallets.

In addition to wooden pallets, plastics pallets are also used. They are often used in the food industry because they are easy to clean due to their smooth surface. The most common material for plastic pallet production is PE, but sometimes PC is used. Pallets made out of PC are stronger than PE pallets. Plastic pallets are especially suitable as multiple-use pallets. They weigh approximately 30 kg. There is no consensus about the number of trips that can be made with a multiple-trip plastic pallet; it varies between 30 and 100 trips.

The shift from one-way pallets to returnable pallets requires major changes in pallet administration and management. The use of multiple-trip pallets requires a pallet pool. A pallet pool is an organization that manages the transit of the pallets between the various users. Standard sizes for pallets are introduced to make the pallet applicable for many users. In Europe, several organizations are active in the management of pallet pools, of which Europool and Chep have the largest market shares.

A third example is the reuse of product components of machinery such as televisions and copy machines. When the product has reached the end of its lifetime, it is collected and taken to a deconstruction facility, where the reusable parts are separated from the other components. The reusable parts are transported to the production facility and can be reused in new televisions and other equipment.

Two types of collection systems are generally used. The first system is completely organized by the producer. In order to make such a system work, the product is sometimes leased and not sold. The advantage is that the producer remains the owner of the product. When the product is not desired anymore by the customer, the producer (and owner) collects the product. This arrangement is typical for large, expensive products such as copy machines. In the second system, the collection is organized by either municipalities that collect electrotechnical products or retail traders that collect old products when selling new products. In The Netherlands, the

latter system exists. When buying new products, a tax is paid to cover the expenses of dismantling the old products.

The success of both systems in terms of product reuse is strongly dependent on the efforts made by the producer to take back old products and reuse product parts. When this responsibility is low, the chances that product parts are actually reused are slim. In these cases, the maximum benefit is that the collection of these products offers potential for material recycling (plastic and metal recycling).

Producers that take product responsibility very seriously include Xerox, which takes back and dismantles its copy machines, and Bosch, which is making efforts to reuse the motors from returned machinery. The lifetime of a motor in electrical equipment is generally much longer than the lifetime of the total product. Reuse of motors is therefore a very good option.

4. ENVIRONMENTAL BENEFITS OF PRODUCT REUSE

The environmental benefits of reuse can be calculated. To do so, the original life cycle needs to be compared with the reuse life cycle. Specifically, all extra processes necessary to make reuse possible need to be analyzed (e.g., separation, cleaning, and transport of used products) and compared to all the processes that do not take place any longer in the reuse life cycle (e.g., material production, material transport, and product manufacturing). The environmental benefit can then be calculated by subtracting the environmental impacts of the obsolete processes from the environmental impact of the additional processes. The environmental benefits of reuse are often quite large since the environmental impact of the separation, cleaning, and transport of used products is often much smaller than the impact of material production, material transport, and product manufacturing. This is emphasized by Table I, which gives an estimate of energy use data for typical processes that occur in the production and use of PET bottles used for packaging of soft drinks. Here, it is assumed that the environmental effects of energy use are a good indicator of the total environmental effects. Table I shows that the dominant processes in terms of energy use are material production and product manufacturing. Since, product reuse leads to reduced material production and product manufacturing, the environmental benefits of reuse are easily explained.

Hekkert *et al.* analyzed a number of improvement options to reduce the environmental impacts of packaging materials. Product reuse was taken into account as an important improvement option. The results of their analyses indicate the potential environmental benefits of product reuse in relation to other options to improve a material cycle, such as material recycling, material substitution, and more

TABLE I

Use of Energy for Typical Processes in the Production and Use of PET Bottles

Process	Energy use (GJ/1000 bottles)
Material production	8.50
Product manufacturing	1.22
Filling of bottle	0.20
Cleaning of bottle	0.45
Transport of bottle	0.10

material-efficient product design. In these analyses, the scope of studying environmental benefits is reduced to studying carbon emissions. In Fig. 2, the results of the analysis for packaging products are depicted.

Figure 2 needs to be read as follows. Each horizontal line represents one specific improvement option. Specifically, it represents the carbon emissions that are saved by replacing a traditional package with an improved packaging concept. Every horizontal line is numbered and the numbers relate to Table II, in which the traditional packaging concept and the improved packaging concept are listed. The length of each line indicates the size of the reduced carbon emissions, and the position of the line in relation to the vertical axis gives information on the life cycle costs associated with this improvement option. Figure 2 shows that the costs of many options have a negative value, which means that over the complete life cycle, costs are saved when the option is implemented.

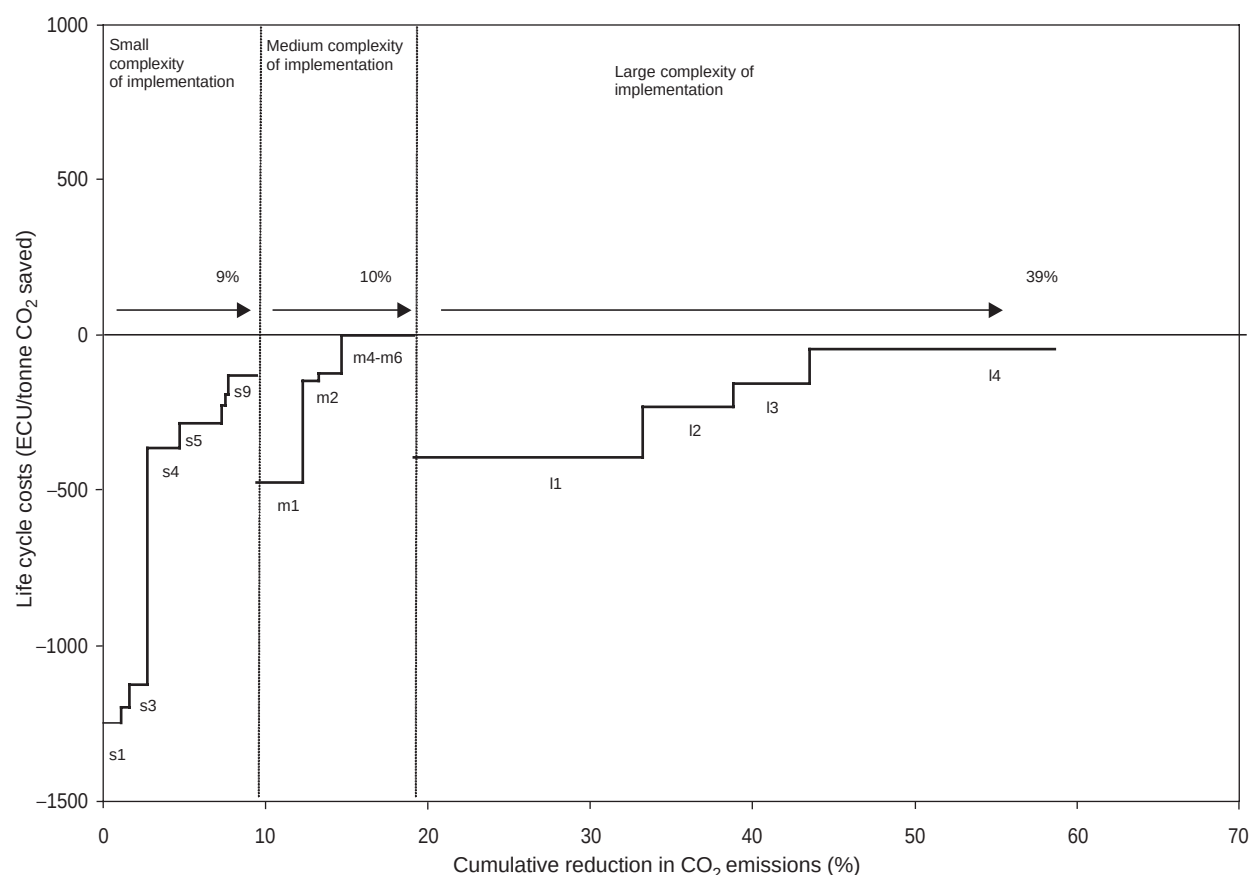


FIGURE 2 A supply curve for the reduction of CO₂ emissions by improved use of materials for primary packaging in Europe. The data refer to Table II.

TABLE II

Potential CO₂ Emission Reduction and Costs of Improved Material Management for Packaging in Western Europe^a

No.	New packaging concept	Old packaging concept	CO ₂ emission reduction (%)	Cost (ECU/tonne CO ₂ saved)
S1	PP film thin	PP film	1.1	-1200
S2	Cardboard box, light	Cardboard box	0.5	-1200
S3	LDPE film thin	LDPE film	1.1	-1100
S4	Honeycomb food can	Steel food can	2.0	-360
S5	Light glass bottle, large	Glass bottle, large	2.4	-280
S6	Glass jar, light	Glass jar	0.2	-280
S7	Steel beverage can, light	Steel beverage can	0.2	-230
S8	Aluminum beverage can, light	Aluminum beverage can	0.2	-190
S9	Light HDPE bottle	HDPE bottle	1.8	-130
M1	PET bottle, one way	Glass bottle, large	5.2	-470
M2	Steel beverage can, light	Aluminum beverage can	1.1	-150
M3	PP cup	PS cup	1.4	-120
M4	PET bottle to be recycled	PET bottle, one way	1.0	0
M5	Recycled HDPE bottle	Light HDPE bottle	2.9	0
M6	Cardboard blister	PVC blister	2.2	0
L1	Pouch	Liquid board	14.1	-390
L2	Glass bottle, small refillable	Glass bottle, small	5.6	-230
L3	Pouch	HDPE bottle	4.7	-160
L4	PET bottle reuse, recycled	PET bottle, one way + PET bottle to be recycled	15.1	-40

^a Abbreviations used: HDPE, high-density polyethylene; LDPE, low-density polyethylene; PET, polyethylene terephthalate; PP, polypropylene; PS, polystyrene; PVC, polyvinyl chloride.

Figure 2 is divided into three parts. The first part indicates the options that can be categorized as easy to implement; it contains mostly options that use a different product design to use lighter or less materials in packaging design. The second part contains options that are more difficult to implement, such as material substitution options. The final part is dedicated to reuse options and product substitution, which are the most difficult options to implement. The figure clearly shows that the total carbon emissions that can be expected from reuse (measure l2 and l4) are very large compared to the other measures. Thus, these results indicate that product reuse can be a very efficient option to reduce the environmental impacts of material life cycles compared to competing measures.

5. IMPLEMENTATION BARRIERS OF PRODUCT REUSE

It was previously mentioned that product reuse can be categorized as a rather difficult measure to improve a

material life cycle. The reason for this is the fact that product reuse requires an additional infrastructure to return the products to the manufacturer or distributor. In this respect, product reuse can be seen as an ultimate form of extended producer responsibility. However, since manufacturers often strive to distinguish themselves from competitors, in every product group a whole range of different product designs can be found. This requires an extremely intense separation and returning system in which all individual products can be traced back to a certain manufacturer. In practice, this is likely not feasible.

Two types of systems do seem feasible. The first system was discussed previously. Here, products are not sold but leased and the manufacturer remains the owner of the product. In this case, the manufacturer knows the location of the product at all times and can take the product back at the end of its lifetime. Typical products for such a system are large, expensive (electronic) equipment for the professional market, such as copy machines. An important drawback of this system is that the speed of technological change in these product types is very high. This

results in a situation in which recycled product parts are often outdated. In these cases, material recycling is the second best option. However, one might question whether for material recycling, returning the product to the owner is preferable to a general collection system. The latter is much easier to organize and can benefit from scale issues.

The second system uses the concept of standardization to create an efficient returning system. In this case, several manufacturers and distributors agree on a standardized product that is manufactured by these parties. Examples are returnable beer and soda bottles in several European countries that do not differ in size, material, or design but only differ in the type of labels used. This system makes it possible for all bottles to be distributed over all manufacturers. This simplifies the setup of the returning system, but it requires the commitment of all major producers and distributors and it decreases the capabilities of the individual producers to develop new competitive product designs. The difficulty in developing this type of returning system is that it requires a large amount of companies (retail organizations, distributors, and manufacturers), the details of the system need to be agreed on by all parties, and much administration is necessary to keep the system going. This leads to transaction costs, which companies prefer to keep as low as possible.

One way to decrease the transaction costs is to make one party responsible for the complete setup and maintenance of the system. Other parties only have to join the system. Such an example is the use of returnable pallets in Europe. In this case, supranational standards have led to fixed pallet sizes, which makes it possible to rotate pallets over distributors throughout Europe. The pallet pools are organized by pool companies that own and maintain the pallets. Detailed tracking systems have been set up to calculate the pallet costs for each company in the pallet pool. The advantage for the distributors is that they do not need to own a large amount of pallets, which saves investment costs and space.

6. CONCLUSION

This article has shown that product reuse is an important method for improved material management, in addition to other methods such as material recycling, improved product design, and improved waste management. Important examples of current reuse practices can be found in the packaging and electronics sectors. The environmental benefit of

product reuse seems quite large due to the fact that major environmental impacts from material production and product manufacturing are largely reduced. The problem with product reuse is that it requires a new and often additional distribution system that requires the involvement of many parties that need to align their actions. Transaction costs are regarded to be high, which can be seen as an important barrier for implementation.

SEE ALSO THE FOLLOWING ARTICLES

Corporate Environmental Strategy • Energy Efficiency and Climate Change • Energy Efficiency, Taxonomic Overview • Industrial Energy Efficiency • Material Efficiency and Energy Use • Recycling of Metals • Recycling of Paper

Further Reading

- Belkom, A. V. (1994). Kunststof pallets in opmars. *Kunststof Rubber* 12.
- Chappin, M. (2003). "Decomposition Analysis of the Dutch Beverage Packaging Waste." Utrecht Univ. Press, Utrecht, The Netherlands.
- Duin, R. V., Hekkert, M. P., de Beer, J., and Kerssemeeckers, M. (1999). "Inventarisatie en Classificatie van EZP opties." Bureau B&G, Emst, The Netherlands.
- Elliott, R. N. (1994). "Carbon Reduction Potential from Recycling in Primary Materials Manufacturing." American Council for an Energy-Efficient Economy, Washington, DC.
- Gielen, D. J. (1999). "Materialising Dematerialisation." Delft University of Technology, Delft, The Netherlands.
- Hekkert, M. P., Joosten, L. A. J., Worrel, E., and Turkenburg, W. C. (2000a). Reduction of CO₂ emissions by management of material and product use: The case of primary packaging. *Resour. Conserv. Recycling* 29, 3–64.
- Hekkert, M. P., Joosten, L. A. J., and Worrel, E. (2000b). Reduction of CO₂ emissions by management of material and product use: The case of transport packaging. *Resour. Conserv. Recycling* 30, 1–27.
- Patel, M. (1999). "Closing Carbon Cycles: Carbon Use of Materials in the Context of Resource Efficiency and Climate Change." Utrecht Univ. Press, Utrecht, The Netherlands.
- Price, L., et al. (1998). Sectoral trends and driving forces of global energy use and greenhouse gas emissions. *Mitigation Adaption Strategies Global Change* 3, 263–391.
- Renia, H. M., and Sikkema, R. (1991). "Houtbijkproducten in Nederland: Een Studie naar de Omvang, Samenstelling, Herkomst en Bestemming van Resthout, Houtafval, oud Hout en Niet-Geoogst Hout." Stichting Bos en Hout, Wageningen, The Netherlands.
- Turkenburg, W. C. (1995). Energy demand and supply options to mitigate greenhouse gas emissions. In "Climate Change Research: Evaluation and Policy Implications." Elsevier, Amsterdam.

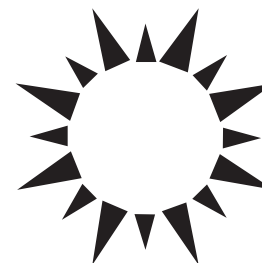
- U.S. Environmental Protection Agency (1998). "Greenhouse Gas Emissions from Management of Selected Materials in Municipal Solid Waste." U.S. Environmental Protection Agency, Washington, DC.
- Worrell, E., *et al.* (1995a). An approach for analysing the potential for material efficiency improvement. *Resour. Conserv. Recycling* 13, 215–232.
- Worrell, E., Meuleman, B., and Blok, K. (1995). Energy savings by efficient application of fertilizer. *Resour. Conserv. Recycling* 13, 233–250.
- Worrell, E., *et al.* (1997). "Potentials and Policy Implications of Energy and Material Efficiency Improvement." United Nations, Department for Policy Coordination and Sustainable Development, New York.

Risk Analysis Applied to Energy Systems

HERBERT INHABER

Risk Concepts

Las Vegas, Nevada, United States



-
1. History of Risk Assessment
 2. Risk Comparisons
 3. Voluntary versus Involuntary Energy Risks
 4. Early and Late Fatalities
 5. Dimensions of Risk Effects
 6. Occupational versus Public Risk
 7. Risk Perception
 8. The Zero-Infinity Dilemma
 9. Storage and Backup
 10. Other Forms of Risk: Climate Change
 11. Air Pollution Risk
 12. Risk Aversion
 13. Uncertainty
 14. Catastrophic versus Noncatastrophic Risk
 15. Comparative Energy Risk
 16. Conclusion

Glossary

aversion A feeling of repugnance toward something with a desire to avoid or turn from it. Risk aversion plays a part in setting environmental standards.

perception Quick, acute, and intuitive cognition. In terms of risk analysis, it has been noted that perception often varies from objective studies.

pollution Environmental contamination with man-made waste. Air pollution is a major source of energy-related risk; water pollution is less so.

radiation Energy radiated in the form of waves or particles. This is a prime concern in terms of nuclear power risk.

risk Possibility of loss or injury.

risk per unit energy or electricity production To put different energy or electricity forms on an equal footing in terms of risk, analysts often divide the risks by a unit energy output, such as a megawatt-year. Other units can and are used. A large fossil fuel or nuclear plant might

have a capacity of 1000 MW. If it operated without interruption for 1 year (highly unlikely), it would produce 1000 MW-years during that time.

uncertainty This may range from a falling short of certainty to an almost complete lack of conviction or knowledge. Risk analyses often contain more uncertainty than policymakers desire.

The production of electricity from any form of primary energy has some, in certain cases substantial, risks to human health and environmental effects. A balanced assessment of different forms of energy requires comparison of their environmental effects with those of their alternatives.

1. HISTORY OF RISK ASSESSMENT

Clearly, our prehistoric ancestors had some appreciation of risk, even in terms of energy. Fire, the main energy source of the time, could destroy unless it was carefully watched and controlled. In a broader context, staying far away from a saber-tooth tiger was safer than having it close at hand. However, until some numerical measure was used, what our ancestors experienced was not what we would call risk analysis. Thus, the phrase implies a type of calculation, from the simplest to the Monte Carlo probabilistic techniques of advanced risk studies.

The Babylonians were apparently among the first to use numbers extensively, often in conjunction with astronomical events. Consequently, it is little surprise that in approximately 3200 BC they apparently had the first risk assessors. These assessors did not deal with health or environmental issues. Although periodic flooding by rivers is now regarded as environmental, the Babylonians believed that these events were created by the gods and not amenable to any

type of quantitative analysis. Issues such as health, death, food supply, and the like were planned by the supernatural, not to be interfered with by humans.

Rather, these forerunners of modern consultants offered advice on other risky decisions, such as deciding where to locate important buildings and when and whom to marry. These areas were clearly not energy related but have some resonance with modern life. Newspaper horoscopes, claimed by their developers to be based on mathematical formulae, are still read by tens of millions of people. In many areas of the world, the dates of major events such as marriages or contract signings are regulated by mathematics. In the United States, it has been claimed that President Ronald Reagan had the dates of some of his decisions set by the same formulae. The purpose of these calculations is to avoid risk—marrying the wrong person or losing money on a business decision.

In the Western world, the industrial revolution generated new diseases or caused the increased spread of old ones. It was realized that some diseases could be prevented or alleviated by human intervention. The scientific study of scrotum cancer in chimney sweeps by Percival Potts in Britain in the late 18th century was perhaps the first epidemiological study. This study has been cited in the first chapter of many books on risk. Since fireplaces were the main source of home heating in Britain at the time, the study was obviously energy related and perhaps the first scientific risk study dealing with energy. Epidemiology has undergirded many energy risk studies since then.

Curiously, however, although detailed in gruesome detail in books by Charles Dickens and others, the terrible diseases and accidents suffered by coal miners of the early 19th century, often children, apparently did not receive the same scientific scrutiny despite the fact that the total loss of health and life by coal miners was much greater than that of chimney sweeps. Chimney sweeps visited the homes of the wealthy, who never saw the suffering coal miners.

This omission still holds a lesson for modern energy risk analysis: The largest sources of risk for a given energy technology are sometimes ignored or glossed over as “common knowledge.” This point will be illustrated later.

By the late 19th century, industrial hygiene, epidemiology, and toxicology were established as scientific areas. However, general application of this research to factories and industrial enterprises had to wait a number of decades as social consciousness

grew. By the 1930s, considerable evidence on the relationship between occupational toxins and human health had accumulated. However, the Great Depression prevented widespread use of these results in regulation. Only certain large firms, such as DuPont, whose social consciousness dated back to its gunpowder manufacturing days, made use of these results.

At that time, it was known that air pollution could cause health problems. Pittsburgh, Pennsylvania, was known as the “smoke city” for the heavy pall it often endured. However, smoke was associated with industrial prowess and jobs, so little was done. During a visit to India in the 1970s, I was given a matchbook showing a factory belching smoke. It is unlikely that such a matchbook would be printed today.

The next era in risk assessment occurred indirectly with the advent of the atomic bomb in 1945. It was quickly recognized that the power of destruction could be harnessed for peaceful purposes, and major governments adopted nuclear power programs. However, the risks associated with producing this form of energy were unlike those calculated in the past. This in turn led to vast research on nuclear risks, much more than for other energy technologies. To date, more has been spent on analysis of nuclear risks than on all other energy sources combined—probably at least an order of magnitude greater.

It was quickly recognized that a major accident at a nuclear reactor could have significant consequences for health, causing cancers and perhaps other diseases. A 1957 study by Brookhaven National Laboratory for an extreme case predicted hundreds of immediate and long-term deaths from radiation. However, defenders of the nuclear industry pointed out that these large numbers were not representative of likely operation.

In the 1970s, drawing on experience from the National Aeronautics and Space Administration, the concept of probabilistic risk analysis (PRA) for complex systems such as nuclear power plants was developed. The methodology was incorporated in the so-called Rasmussen Report of 1974–1975, which showed that the average risk to health from nuclear reactors was extremely small. As in many phases of life, the average is considerably different from extreme events or conditions.

Since that time, dozens of PRAs have been performed for nuclear reactors (as well as for chemical and other large industrial facilities). In addition, because energy has assumed such a major role in modern life—the famed economist Julian Simon once called it the “master resource”—thousands of studies on various aspects of energy risk

have been completed. It would take far too much space to even list, let alone discuss, them here.

2. RISK COMPARISONS

Risk comparisons are used extensively in risk analyses for a simple reason: Lay audiences often have difficulty understanding the complex terminology and mathematics of many energy risk studies. This has been shown in both psychological studies and public meetings on energy. In this section, the reasons for risk comparisons are discussed in a general way. Later, specific comparisons are evaluated.

Regulations in the United States and many other industrialized countries require public meetings to be held to discuss new energy projects. When the subject of risk arises, observers have noted that there is little, if any, substantive contribution from the public. When roentgens, rems, and parts per billion are mentioned by speakers, the relatively low level of scientific understanding by the public becomes apparent. To remedy this, risk comparisons are often used. These comparisons can be divided into two categories: (i) comparisons between different substances and activities and (ii) comparisons between different technologies.

The first category cuts across different aspects of life and exposure. For example, a widely used comparison devised by Richard Wilson of Harvard University states that each of the following activities presents the same risk, measured as the increased risk of death within 1 year by 1 in 1 million. This means, for example, that cycling the specified distance will increase that person's chances of death within a year by one in a million, a very small chance:

- Smoking 1.4 cigarettes
- Riding 10 miles by bicycle
- Eating 40 tablespoons of peanut butter
- Drinking 30 12-ounce cans of diet soda containing saccharine
- Living within 5 miles of a nuclear reactor for 50 years

It is evident that calculating each of these risks contains a wide variety of data and assumptions. For example, nonsmokers (approximately three-fourths of the population) who smoke one or two cigarettes almost certainly will not develop lung cancer as a result. The estimate of one chance in a million of death is based on smokers. Peanut butter by itself will not cause death; any aflatoxin it contains might.

Some risk experts have objected to these types of comparisons, claiming that the activities specified have little or nothing to do with each other. Peanut butter has no known connection to nuclear power. They state that comparing the risks of different forms of energy is more illuminating to the public than contrasting the risks of smoking with nuclear power. The problem is reduced to ensuring that the entire fuel or energy cycle for each energy technology is used to allow for a fair and reasonable comparison.

3. VOLUNTARY VERSUS INVOLUNTARY ENERGY RISKS

Chauncey Starr was one of the first to quantify the difference between voluntary and involuntary risks. It is self-evident that risk, at least per hour of activity, is greater for hang-gliding and mountain climbing than for driving an automobile or sitting behind a desk. Starr estimated that voluntary activities were between two and three orders of magnitude more risky than involuntary ones, depending on the precise activities considered.

How does this translate to energy? With some difficulty. The problem is that the line between voluntary and involuntary activities, which is clear in other contexts, is often unclear in the energy situation. For example, consider a coal miner in Wales or China. He may be descended from a long line of miners and is expected to go into the mines when he reaches working age. In principle, coal mining, with its known high risks, should be a voluntary activity. Those who do not wish to accept its risks are able to take up other less risky occupations. However, in the example cited, it is not completely clear that it is voluntary.

In terms of public risk, again the line is often indistinct. Certain areas of the nation have higher levels of air pollution, mostly due to energy activities, than others. Los Angeles has its infamous smog. Again, in principle, residents who know about these air pollution risks and are concerned about them can move to other more bucolic areas of the nation, making this air pollution exposure voluntary. Some do. However, many are confined by employment and family ties to the polluted region, suggesting that the risk is more involuntary.

The question becomes more acute with regard to nuclear power, clearly the most highly publicized form of energy risk. When a new nuclear reactor is to be built, residents have the option of moving away if

their concerns about risks are great. No doubt some have done so in the past, although there do not seem to be sociological studies indicating the percentage of those who depart. It is likely to be small for the reasons mentioned previously as well as others, such as difficulty in selling a home and sheer inertia.

Thus, although using the dichotomy of voluntary versus involuntary risks can be useful in other aspects of risk analysis, it is difficult to employ in the energy context.

4. *EARLY AND LATE FATALITIES*

Deaths from energy technologies can be characterized as early or late. The division applies primarily to nuclear power. For other energy technologies, such as coal- or gas-fired electricity, deaths occur almost immediately after an accident or air pollution incident. For example, construction and mining accidents, which are shown elsewhere in this article to be a substantial part of some energy technology risks, produce deaths either immediately or within a few days. The effects of air pollution from fossil fuels are often calculated by the increase in deaths after a pollution alert. When a dam fails, deaths are immediate.

Nuclear power is different in the sense that if a major accident occurs, the release of radiation could cause deaths from cancers in three or four decades. The Three Mile Island (TMI) accident in 1979 did not produce enough radiation to generate this effect. Although the radioactive release from Chernobyl in 1986 was substantially greater than that of TMI, the World Health Organization and International Atomic Energy Agency calculate that the late fatalities will be small. However, they will be nonzero.

It is evident that both early and late fatalities from a nuclear reactor accident should be combined. The Rasmussen Report was criticized for not including late fatalities with its estimates of early deaths from a reactor accident.

5. *DIMENSIONS OF RISK EFFECTS*

Decision makers want one or two numbers in a risk comparison. Scientists would prefer to have many dimensions of a subject quantified. The tug-of-war between the two viewpoints has persisted in the past and is likely to continue indefinitely.

These generalities can be made concrete by specific examples. Inhaber determined that the risks

of the energy systems he considered could be divided into two categories: public and occupational. This division is described by some as involuntary and voluntary, respectively. As noted earlier, the division is not so clear-cut. He then noted that some risk estimates were in deaths and others were in nonfatal diseases and accidents. A number of studies estimated the loss of life expectancy for fatalities, accidents, and diseases due to energy production. The data for both occupational and public effects could then be combined into man-days lost.

The next step was to provide a common denominator for the amount of energy or electricity produced by the 11 systems he considered. He chose a megawatt-year, but in principle any multiple of this could have been used.

Since he was aiming his research at decision makers, he combined the occupational and public risks, both in man-days lost per megawatt-year, into a total risk. For this, he was criticized by some for oversimplification and combining two dissimilar entities.

One of the valid criticisms of risk analyses of that era was that they did not take into account the ages of victims of accidents or diseases. Clearly, someone of age 80, perhaps already in poor health, has fewer days lost if felled by an air pollution spike than a vigorous 21-year-old killed in a mine accident. Because governments have an interest in the efficacy of health interventions, health scientists have devised the "quality days of life lost," which takes into account the previous points concerning relative ages of victims of energy risk as well as their state of health. This measure provides a more accurate view of energy risk effects.

Researchers at Clark University in Massachusetts took another approach. They attempted to characterize technologies on a scale that encompassed 12 descriptors, including social, physical, and biological aspects of risk: intentionality, spatial extent, concentration, persistence, recurrence time, population at risk, delay of consequences, annual mortality, maximum potentially killed, effects across generations, maximum potential nonhuman mortality, and experienced nonhuman mortality. Although some of the descriptors are self-explanatory, others would require a lengthy discussion of their assumptions.

However, the problem still remains: Is there a rational and defensible method of combining 2 or 12 measures, or should they be presented separately? For example, if technology A has a lower risk than technology B on 3 of 12 measures but also has a higher risk than technology C on 5, how do we rank

A, B, and C? It is unlikely that decision makers would have the skill or patience to sort through all the possible permutations and combinations. For this reason, at least some risk researchers will continue to combine different measures of risk into one or two easily understood numbers.

6. OCCUPATIONAL VERSUS PUBLIC RISK

In the energy risk studies discussed elsewhere in this article, risk to workers tends to be higher than risk to the public. In a completely rational world, emphasis should be on reducing worker risk, with less concern about public risk reduction.

Al Hoskin of the U.S. National Safety Council evaluated the relative risks to workers and the public in Superfund (abandoned waste) sites. He found that the risk to workers for the cleanup was higher than the existing risk to the public.

In reality, developed nations place the emphasis on public risk. As noted later, this may be due to the media's concentration on disasters, real or potential, affecting the public. Occupational risk to individual workers in the United States tends to receive little attention, except by the agencies devoted to the problem or when a disaster occurs, as in the case of the Pennsylvania coal miners trapped below the surface in water in 2002. The miners were rescued with no loss of life. At the same time, news came of major coal mine disasters in China, where scores were killed. Little attention was paid to these incidents.

7. RISK PERCEPTION

The following is one of the oldest questions in philosophy: If a tree falls in a forest and nobody hears it, does it make a sound? Like many philosophical issues, there is no simple answer. The analogous question in risk analysis is, If a study shows that chemical or toxin has a high (or low) risk, and this study is either unknown to the public or ignored by it, what then is the risk?

As Slovic, Fischhoff, and Lichtenstein (SFL), three of the founders of risk perception studies, have noted, subjective judgments, by both experts and lay people, are a major component of any risk assessment. Expert judgments are often obscured in

impenetrable mathematics. Judgments of lay people are considerably easier to detect.

In a classic study that went from the social science literature to popular magazines, SFL noted the divergent risk estimates of four groups: members of the League of Women Voters (LWV), college students in Oregon, a business and professional club, and risk experts. SFL considered 30 sources of risk, from nuclear power to lawn mowers and vaccinations. Many of these activities rarely appear on risk lists, but they all produce some hazards. Only a few of the categories related to energy risk.

The two largest divergences between estimates of the lay group and experts were in nuclear power and X-rays. The experts rated nuclear energy as 20th riskiest of the 30 activities, the LWV and college students rated it highest, and the business and professional people rated it 8th. X-rays were rated approximately 20th in riskiness by the laypeople but 7th by the experts. The rankings of many of the other risk sources were approximately the same for laypeople and experts.

Although there are many possible explanations for why the estimates varied so much, the variation results from the benefits received from the two activities. X-rays are viewed as having considerable benefits in medical diagnosis that outweigh the risks of radiation. The benefit of nuclear power is electricity generation, but there are other ways of producing this energy form, such as coal, gas, and oil. In the United States, approximately 80% of electricity is nonnuclear. (Admittedly, there are methods other of detecting internal diseases besides X-rays, such as nuclear magnetic resonance, but these were not included on the SFL list.) Thus, the benefits and risks of nuclear energy must be compared not to those of swimming or lawn mowers but to those of other ways of generating electricity.

One aspect of the SFL study that deserves mention is the ranking of technologies as opposed to concentration on absolute values of risk, either real or perceived. Because the values in their studies ranged over many orders of magnitude, it was clearer to rank technologies rather than to discuss the absolute values.

Experts measure risk as X deaths or diseases per activity, but risk is perceived according to SFL by whether it is voluntary, controllable, certain to be fatal, catastrophic, immediately manifested, and unfamiliar. Some of these are discussed elsewhere in this article.

Most of the perception aspects of risk can be described by two factors: dread (related to aspects

such as certain to be fatal and catastrophic) and how unknown the risk was (related to involuntariness and newness). On these scales, nuclear power had relatively high scores for both factors, even among the experts.

How are risk perceptions formed? The media clearly plays a major role. For example, pesticides ranked approximately eighth in the list of 30 activities noted previously. Some years ago, concerns about the pesticide Alar were expressed on a widely watched television show. Would a repeat of the poll have shown higher perceived risk for pesticides? Risk assessors sometimes talk about “risk of the week” in reference to the media publicizing a seemingly new hazard on a regular basis. On the other hand, editors of newspapers and television news shows would respond that they are merely reporting on what interests the public.

8. THE ZERO-INFINITY DILEMMA

Students of mathematics know that multiplying zero by infinity produces an indeterminate value—that is, no number that can be identified. What relationship does this have to energy risk?

The problem arises in only one energy technology—nuclear. It is acknowledged by the nuclear industry that in the most extreme conditions, a reactor accident can emit considerable radiation, thus producing deaths from disease. The number of deaths is clearly not infinite but can be high.

The industry points out that, at least according to probabilistic risk analyses, the chance of such an accident occurring is extremely small, perhaps one in tens or hundreds of thousands of years of reactor operation. The chance is not zero, but very small. Thus the zero-infinity question: If one multiplies a very small number by an enormous one, what is the product?

Simple calculation shows that the result is a small risk. However, the estimate—and both the “zero” and “infinity” are no more than estimates—is conditioned more by risk perception than by mathematical analysis. Many people think of the risks of nuclear power primarily in terms of the quasi-infinite number. They tend to ignore or disregard the other factor in the product—the estimate of likelihood. Thus, the risks of nuclear power may be, in some minds at least, overestimated.

Curiously, the catastrophic risks of a renewable resource, hydroelectricity, do not seem to be weighed in the same scale. The vast majority of hydro dams are as safe as any structure on Earth. However, there

have been major failures over the years of some dams, in some cases causing hundreds of deaths in the valleys below. The zero-infinity dilemma does not seem to be applied to this energy form, although the deaths it has caused are real, not hypothetical.

9. STORAGE AND BACKUP

No energy system operates at 100% capacity all the time. Even in the best run systems, parts have to be adjusted and replaced, and unforeseen accidents can occur. On average, nuclear and fossil fuel systems generally operate in the 80–90% capacity range. That is, approximately 10–20% of the time a given plant in this group is unavailable for refueling, maintenance, or other reasons.

The vast majority of these shutdowns are planned. If fossil fuel and nuclear plants generally stopped on a random or arbitrary basis (these types of shutdowns can occur but are comparatively rare), it would be extremely difficult to have a reliable electricity grid.

The situation is different for at least some renewables. They require some combination of energy storage and backup—the supply of conventional energy when the sun does not shine and the wind does not blow. Consumers generally demand highly reliable electricity service. The proof of this is the torrent of phone calls that a utility receives from irate customers when their power is temporarily cut off due to weather-related accidents.

For hydroelectricity, categorized as a renewable energy source by the U.S. Energy Information Administration and other agencies, the entire system is storage. Thus, no further storage or backup is necessary. Using wood to produce electricity also does not require backup since enough wood can be gathered to provide as reliable fuel supply as fossil fuels. Transportation using corn-based fuels also does not require backup since the sun’s energy is stored in the corn.

However, the variability of the sun and wind require some combination of storage and/or backup. Some quotations from the literature on solar and wind make the point:

“Natural gas serves as the backup heat when the solar energy is unavailable or insufficient.”

“A global solar-energy system without fossil or nuclear backup would also be dangerously vulnerable to drops in solar radiation from volcanic events such as the 1883 eruption of Krakatoa, which caused

widespread crop failure during the ‘year without a summer’ that followed.”

“Luz (solar thermal) units are hybrid systems with natural gas firing capacity added as backup. ... PV systems (and wind systems as well) can be used in hybrid configurations with natural gas-based backup power, as in the case of solar thermal units.”

“A wind-driven generator to convert wind energy to electrical energy, backup diesel generators (for when there is no wind). ...”

“The principal disadvantage of solar is the dependence on variable sunshine hours and the resulting short capacity utilization as well as the need for energy storage and backup systems.”

Thus, the question is, what fraction of backup and/or storage should be required for an intermittent system? Estimates have varied up to 20% of the original system for backup. This implies that up to 20% of the risk associated with conventional energy systems is attributable to solar or wind systems. Other estimates of storage and backup are lower.

Electricity cannot be stored, except in expensive batteries impractical for an electrical grid. It is true that over a sufficiently large area, the wind is always blowing somewhere, suggesting that wind power is continuous, not intermittent. However, no technology can instantaneously transfer wind power from North Dakota to Florida, for example. If and when such technology is developed, the need for backup and storage will disappear.

10. OTHER FORMS OF RISK: CLIMATE CHANGE

The risks discussed in this article are primarily to health, in the form of accidents or diseases. However, there is another less apparent form of risk—that of climate change. The greenhouse effect refers to the effect of certain trace gases in the earth’s atmosphere so that long-wave radiation, such as heat from the earth’s surface, is trapped. Much of the greenhouse effect is due to carbon dioxide. An increase from 280 to 350 parts per million has occurred since the beginning of the industrial revolution.

If the connection between increased concentrations of carbon dioxide and rising world temperatures is valid, it has been stated by many climatologists that the result will be more hurricanes, tornadoes, droughts, and other extreme events. In this sense, risks to humans will be greater, if the chain of assumptions is valid.

Renewables and nuclear electricity produce no carbon dioxide directly. However, there are indirect effects, albeit much less than the direct combustion of fossil fuels to produce electricity. The electric car is sometimes held up as an example of a system emitting no carbon dioxide, which is correct in a direct sense. However, approximately 70% of the electricity used to charge its batteries in the United States derives from fossil fuels. The carbon dioxide is emitted at the plants, not on the street.

The global climatic effect of increasing CO₂ levels is a factor in the comparison of different forms of energy. Worldwide emissions of CO₂ from burning fossil fuels total approximately 25 billion tonnes per year. Approximately 38% is from coal and approximately 43% from oil. Every 1000-MWe power station running on black coal produces CO₂ emissions of approximately 7 million tonnes per year. If brown coal (a less energy-intensive type) is used, the amount is much greater.

Figure 1 shows the estimated production of CO₂, both direct and indirect, for different energy sources. As noted elsewhere in this article, there are often wide variations in estimates of energy risk and their subcomponents. Also, the significant figures listed later for some of the data points should be treated with caution.

Coal has by far the highest values, and almost all reports on the subject confirm this result. The maximum direct production of CO₂ for coal is approximately 1000 g/kWh. Indirect production (coal mining, equipment used in mining and electricity production, etc.) is approximately 300 g/kWh, for a total of 1300 g/kWh. The minimum of the range is approximately 1000 g/kWh.

The range for gas is approximately half that of coal. The other four energy sources do not have

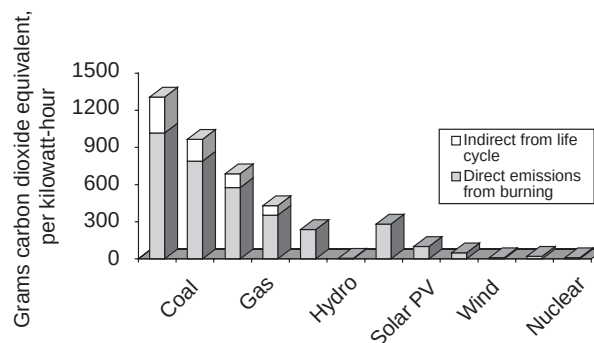


FIGURE 1 Greenhouse gas emissions from electricity production. Modified from the International Atomic Energy Agency (2000).

direct emissions. Hydroelectricity has values ranging from 4 to 240 g/kWh. One reason for the wide range of almost two orders of magnitude may be that the lower values refer to solid-earth hydro dams. Many hydro dams in developed countries are exemplified by the Hoover Dam, built in the 1930s near Las Vegas, Nevada. Much of the dam is concrete and steel, the making of which generates CO₂.

Earthen dams require soil and rock from the surrounding countryside, the transport of which generates only small amounts of CO₂. To give a representative value for hydro, indirect CO₂ production from dams built during approximately the past generation would have to be averaged.

Solar photovoltaic systems have values ranging from 100 to 280 g/kWh primarily because of the large amount of materials used per unit energy output in their construction. The values are substantially higher than those for hydro, by far the largest renewable energy system used throughout the world to generate electricity.

Wind energy has values from 10 to 50 g/kWh. Its lower range is thus higher than that of hydro, but its upper range is lower.

Nuclear has a range of 9–20 g/kWh. It is possible that the variability in nuclear data is lower than that of most other energy systems because so many studies have been performed in this area. Although nuclear, like the three other renewable systems described in Fig. 1, emits no direct CO₂, substantial amounts of metals such as steel and zirconium are used in construction of reactors. Also, given the extremely thick walls of most Western-built reactors, large volumes of concrete are used. In many countries (excluding those using the Canadian CANDU system), enrichment of the uranium for fuel rods employs fossil fuels for the required electricity.

These sources of CO₂ production are divided by the large amounts of electricity produced by typical reactors. Thus, nuclear has CO₂ values comparable to those of wind and considerably better than those of the other systems, both renewable and fossil fuel, considered in Fig. 1.

In summary, coal and gas have substantially higher emissions of CO₂ compared to renewables and nuclear. Of the renewables, solar photovoltaic seems to be highest and wind the lowest. Nuclear has values comparable to those of wind. There are considerable uncertainties regarding all the values.

11. AIR POLLUTION RISK

The largest source of risk to the public associated with fossil fuels has been air pollution effects. These effects can range from death to asthma attacks and other breathing difficulties.

However, a report in a 2002 issue of the Canadian *National Post*, quoting an article in the journal *Nature*, suggests that the estimate of 5000 Canadians (mostly elderly) dying each year from air pollution is too high by at least 1200 deaths and very likely by 3000–4000. This implies that air pollution deaths may be overestimated by approximately 30% and possibly by 150–400%. It is clearly too soon to determine the validity of these claims. Most comparative risk studies suggest that fossil fuels have higher risks per unit energy output than renewables and nuclear. If the study is validated, it would make fossil fuels closer on an absolute scale to renewables and nuclear per unit energy output. However, the relative position of these former fuels compared to their competitors probably would not change.

12. RISK AVERSION

There is a wide range of attitudes toward risk, as shown by surveys and other indirect devices. Risk-seeking individuals drink heavily, drive fast sports cars with unbuckled seat belts, smoke, sky-dive, climb mountains, and engage in other well-known risky behaviors. Risk-averse individuals read the newspapers avidly for the next food or activity to avoid, try vegetarian diets, wash their vegetables and fruits twice, bundle up heavily for cold weather, eat an apple every day, take their medicine when prescribed, move away from nuclear reactors if one is built nearby, stay away from people with colds, and try to elude all risks.

Some individuals are risk seeking for certain hazards and risk averse for others (e.g., smoking while practicing vegetarianism). Most people are somewhere in the middle, steering clear of the obvious risks but not having enough time to study and avoid all hazards.

Policymakers must account for both ends of this spectrum. If most people were risk seeking, there would be little need for any government regulation of energy or any other activity. If most were risk averse, virtually all economic activity would halt. Ironically, this would probably increase societal risk dramatically as society retreated to preindustrial age conditions.

Although many risk calculations are publicly available, there is little evidence that those who are risk averse and risk seeking make use of these data. Risk-averse people are unlikely to take more risks if they are told that they are concerned about infinitesimal hazards. Conversely, risk seekers know that climbing mountains is risky without information from a study. The inclination to join either group is a function of personality, not the results of risk assessments that people have read.

The risk averse and risk seekers seem to coexist happily in most situations. One situation in which they come together is when a waste site is being proposed nearby. Although risk estimates may indicate that the hazards of these sites are extremely small when regulations are satisfied, risk-averse and even risk-seeking citizens generally show strong concerns at siting meetings. As a result, most of these proposed sites are either abandoned or take many years to complete.

13. UNCERTAINTY

To paraphrase the poet, risk analysis, thy name is uncertainty. From the first energy risk analysis, uncertainty has made the results less convincing than they might have been. Some energy risk analyses have not discussed their uncertainties at all. Most others have done so, but sometimes the final results do not reflect the detailed uncertainty calculations in appendices. For example, in the Rasmussen Report, uncertainties were discussed in great detail in the body of the work. However, in the executive summary, which was read by far more people than were the elaborate calculations, uncertainties were downplayed. This led to criticisms.

On the other hand, the Slovic study on risk perception apparently did not show uncertainties at all. That is, if one or another group were ranking nuclear power or lawn mowers in terms of their risk, there would certainly be a statistical distribution. These distributions were never shown in the many reprints and reworkings of the Slovic study. Curiously, the work was not criticized for this omission, although it dealt with many of the risk issues that the Rasmussen Report and many other analyses dealt with.

It would take considerable space to discuss all the uncertainties associated with energy risk analyses, from data to statistical distributions and variations in assumptions. It is sufficient to state here that one of the main charges leveled at risk analyses is that the

uncertainties are sometimes glossed over or not discussed at all.

14. CATASTROPHIC VERSUS NONCATASTROPHIC RISK

Many members of the public tend to think of energy risk in terms of catastrophes, those that kill or injure many at a time. However, at least for industrialized countries, most energy risk is noncatastrophic—that is, occurring to perhaps less than five people at a time.

Nonetheless, there have been catastrophes associated with energy production. Some of the major ones from 1977 to 1993 are shown in Table I. (A full listing to the present would take too much space.) Even for risk experts, many of the listings are likely to be unknown or, at best, barely recalled. In terms of coverage by Western news media, the reports tend to appear somewhere on page 13 of newspapers, with perhaps a 5-s news clip on television. Because many of them occur in Third World countries, the disasters are rarely followed up by stories explaining why they took place.

A number of comments can be made about Table I. First, many major accidents occurred prior to 1977. For example, the Vajont dam accident in Italy killed 2600–3000; The St. Francis dam failure in the United States in 1926 killed more than 400. A complete table would occupy many pages.

Second, many of the accidents took place in Third World countries, as noted previously. Because safety and environmental controls are usually better in industrialized nations, this is to be expected.

Third, although the Chernobyl nuclear accident undoubtedly received more media coverage than all the other accidents combined, it is clear from the table that the vast majority of catastrophic deaths were due to two major energy systems: fossil fuels (mostly gas and coal) and hydroelectricity. Because of their small size, renewables, with the notable exception of hydroelectricity, are unlikely to have catastrophic accidents. The largest recorded in the table was the Indian Machhu II dam failure in 1979.

Fourth, the question arises, can catastrophic and noncatastrophic risk be added in a reasonable way to produce an overall risk value for each technology? There is no simple answer to this question. However, it should be noted that catastrophic risk is small compared to ordinary risks for virtually all energy technologies. Thus, whether or not the catastrophic aspects are added is unlikely to change the comparative energy risk rankings.

TABLE I

Energy-Related Accidents, 1977–1993^a

Location	Year	No. killed	Comment
Machhu II, India	1979	2500	Hydroelectric dam failure
Hirakud, India	1980	1000	Hydroelectric dam failure
Ortuella, Spain	1980	70	Gas explosion
Donbass, Ukraine	1980	68	Coal mine methane explosion
Israel	1982	89	Gas explosion
Guavio, Colombia	1983	160	Hydroelectric dam failure
Nile River, Egypt	1983	317	LPG explosion
Cubatao, Brazil	1984	508	Oil fire
Mexico City	1984	498	LPG explosion
Tbilisi, Russia	1984	100	Gas explosion
Northern Taiwan	1984	314	Three coal mine accidents
Chernobyl, Ukraine	1986	31 +	Nuclear reactor accident
Piper Alpha, North Sea	1988	167	Explosion of offshore oil platform
Asha-ufa, Siberia	1989	600	LPG pipeline leak and fire
Dobrnja, Yugoslavia	1990	178	Coal mine
Hongton, Shanxi, China	1991	147	Coal mine
Belci, Romania	1991	116	Hydroelectric dam failure
Kozlu, Turkey	1992	272	Coal mine methane explosion
Cuenca, Ecuador	1993	200	Coal mine

^aFrom the Uranium Information Centre (2000).

15. COMPARATIVE ENERGY RISK

15.1 Can We Learn from Comparative Energy Risk?

In principle, comparing the risks of different forms of energy can guide both the public and decision makers to more valid decisions and attitudes. In practice, the conclusions are not so simple to draw.

Some observers, such as Andrew Stirling, have suggested that there is so much overlap between the results of energy risk analyses that few if any conclusions can be drawn. Figure 2 shows the results of approximately 30 energy risk analyses that he compiled. Certainly, the wide overlap between each technology suggests considerable uncertainty in conclusions.

Nonetheless, some crude inferences can be drawn even from Fig. 2. The fossil fuels generally have maximum values substantially higher than those of nuclear and the three renewables shown. Hydroelectricity is in the middle range.

In 1996, the Health Physics Society suggested that below a lifetime dose of 100 mSv (10 rem) (a unit of radioactive dose), “risks of health effects are either too small to be observed or are nonexistent.” If this is correct, it is likely that the maximum value assigned

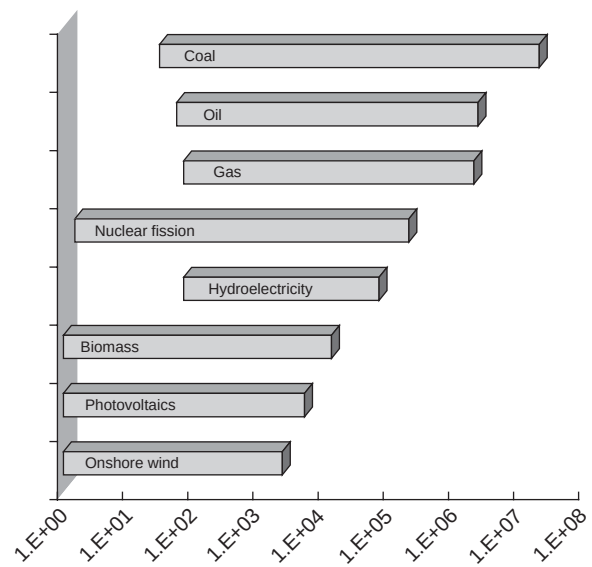


FIGURE 2 Total risk costs for selected energy systems in 1995 U.S. dollars per gigawatt-hour. Modified from Stirling (1999), illustrating ambiguities in the environmental ranking of contending options.

to nuclear fission in Fig. 2 would be substantially lower than the corresponding value for hydroelectricity, making it comparable to the renewables.

The problem with figures of this type is that they compare vastly different studies, with often strongly disparate assumptions. If similar studies are compared, the overlap may be substantially less. An analogy may be drawn about the heights of men and women. On average, men are taller than women. However, in a random group of perhaps 100 of each gender, the tallest woman will almost certainly be taller than the shortest man. The overall conclusion that men are taller than women is not disproved by this observation. Applying this to Fig. 2, the overlap may not be as great as it appears if like studies are compared.

15.2 Being Comprehensive

If comparisons are to be made between energy systems, great care must be taken to include all the sources of risk, from mining of the raw materials to disposition of waste, if any. This generalization can be difficult to implement in practice. Should the risks associated with building the machinery to do the mining be included? Should storage and backup, if necessary, be included as well? There has been considerable debate in the literature on these and other topics of comprehensiveness, without final resolution.

One way of solving at least some of these difficulties is the use economic input-output (I-O) tables, which show how different inputs to a system result in outputs. The risks associated with the inputs and outputs can, in principle, be calculated. Although the I-O system has been used by a number of investigators, it has not been generally adopted in

the risk arena. Most investigators use a system approximately as shown in Fig. 3.

The top of Fig. 3 refers to occupational risks and the bottom to public risks. Clearly, there is considerable simplification in such a figure. For example, waste disposition, one of the greatest public concerns about nuclear power generation, is not shown. At least a rough estimate, given the fact that nuclear waste can pose some risks for thousands of years, is included in many studies of this technology.

Nonetheless, a figure of this type, with its defects, is useful. It shows decision makers, who rarely have time to sort through risk analyses of many volumes, the conceptual basis of comparative risk analysis.

15.3 The Hauptmanns Study

In 1991, Hauptmanns in Germany attempted to summarize more than a decade of major energy risk studies. He considered a wide variety of sources. Rather than develop a statistical distribution, he tabulated the maximum and minimum for 15 energy and electricity sources. Some of them, such as nuclear fusion, are mostly speculative and are not shown in Fig. 4.

In the interests of brevity, only occupational risks are discussed. Hauptmanns also estimated public risks for these technologies. Only fatalities are indicated in Fig. 4. As noted elsewhere in this article, an overall measure of occupational or public risk would also include injuries and diseases attributable to energy systems.

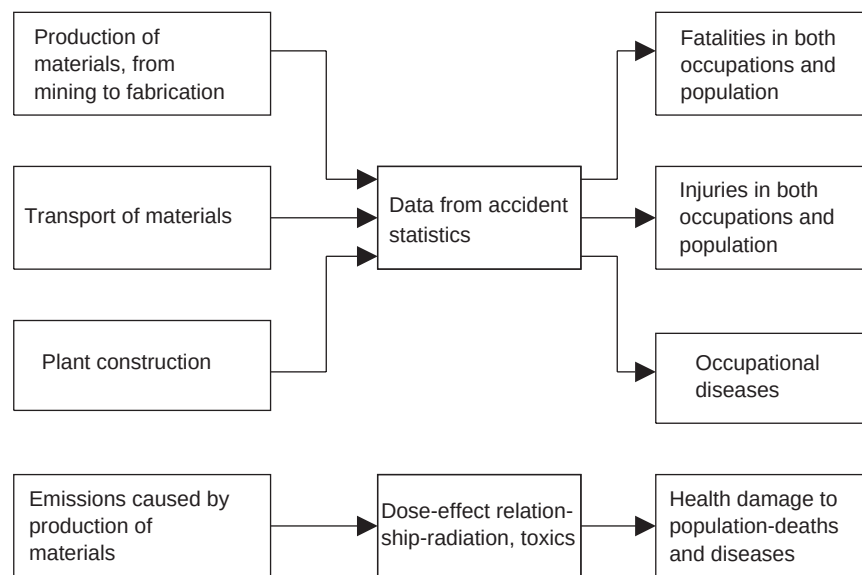


FIGURE 3 Sources of overall risk for energy systems. Modified from Hauptmanns and Werner (1991).

Although there is clearly overlap for many of the ranges shown, it is somewhat less than indicated by Stirling. The highest value seems to be for methanol, beyond the range of the 10 other systems. The fast

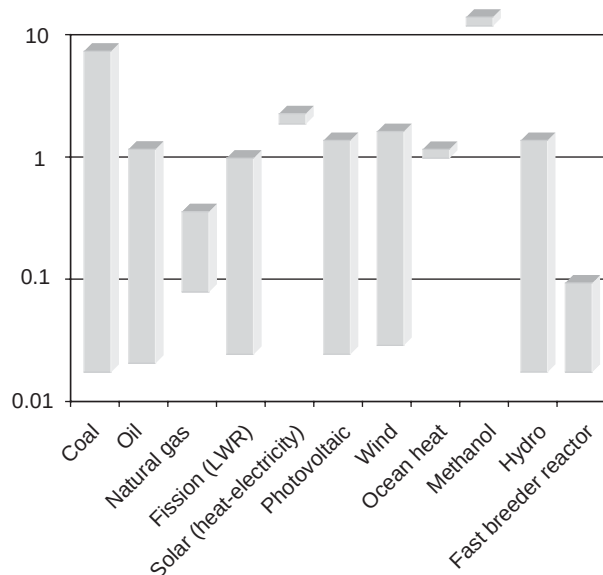


FIGURE 4 Employee (occupational) fatalities per 2.2×10^{16} J for 11 energy systems. A similar graph can be prepared for public risks. Modified from Hauptmanns and Werner (1991).

breeder reactor apparently has the lowest value. However, the upper part of its range overlaps with that of a number of other systems, such as coal, oil, solar photovoltaic, and wind.

The next highest value is for coal, at least judging from its maximum. Its minimum, however, is among the lowest. This may be attributable to assuming so-called strip mining, for which occupational risks are substantially lower than those for underground mining. A fair estimate of coal occupational risks would allocate appropriate percentages to underground and surface mining.

15.4 The ExternE Project

After the Hauptmanns effort, the European Union, through the ExternE project, and the International Atomic Energy Agency, through its DECADES project, supported the largest schemes ever to estimate the relative risk of different energy and electricity systems. Taken as a whole, they are the most comprehensive set of energy risk studies to date.

Results are mostly applicable to Western Europe and, to some extent, the United States. Although evaluating the risks of energy in Third World countries would be useful as they industrialize,

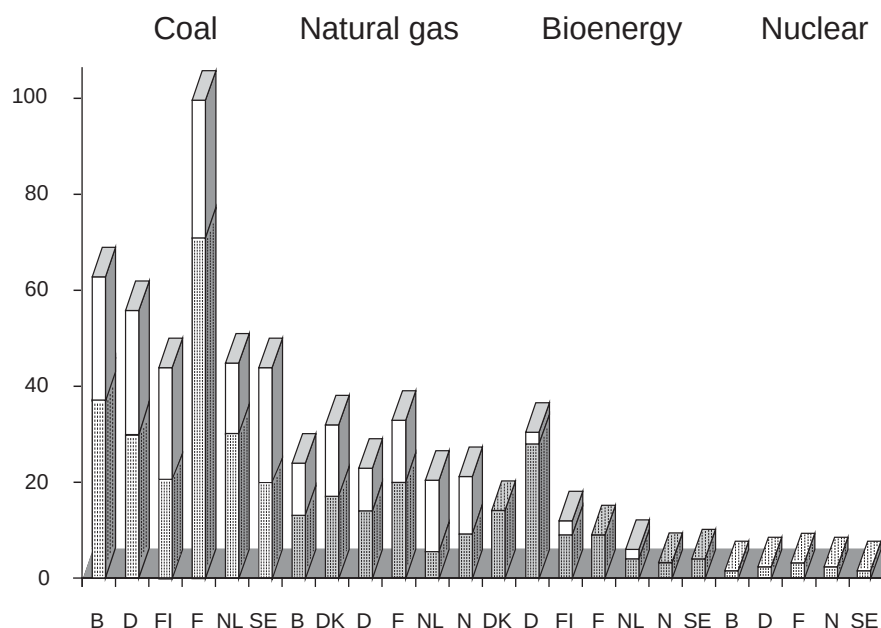


FIGURE 5 External costs of the entire fuel cycle, including global warming, in mECU per kilowatt-hour. B, Belgium; D, Germany; A, Austria; F, France; NL, The Netherlands; SE, Sweden; DK, Denmark; FI, Finland; N, Norway. The shaded portions of each column show the total environmental costs, including global warming, at a relatively low cost per ton of carbon dioxide emitted; the unshaded portions show the total costs at a higher value per ton of carbon dioxide emitted. Modified from Starfelt (1998).

there has been comparatively little written on the subject.

Starfelt attempted to draw general conclusions for the ExternE work. Results are shown in Fig. 5. He considered four systems: coal, natural gas, bioenergy, and nuclear. Different studies were performed for a variety of European Union countries. All costs, risks to human health and the environment, were put on a fiscal scale. The European Community Unit (ECU, commonly known as the euro) was at a level of approximately \$1.20 in December 2003.

Note that this is similar to the calculations of Stirling noted previously. In effect, human risks per assigned a monetary value, requiring an estimate of how much a human life is worth. Although this estimate is required for the ExternE calculations, it is so buried in the voluminous calculations that no easy extraction is available. The reduction of all costs to money also implies estimates of the financial effects of climate change, which are not completely known. Thus, the ExternE results require more assumptions than the standard risk calculations for energy, which have already been criticized for having too many assumptions.

In Fig. 5, the shaded areas show the external costs, including global warming at an estimated cost of 19 ECU per ton of CO₂. Note that this is not a direct measure of potential economic costs of global warming but rather an estimate of the price that firms might pay to emit one extra ton of CO₂ in the future. The columns including the blank areas assume a cost of 46 ECU per ton of CO₂.

Despite substantial variations, coal had by far the greatest external costs, regardless of the estimated cost per ton of CO₂. French coal had the highest values, perhaps because coal from this nation is riskier to mine than elsewhere. The added cost of 46 ECU/ton made coal substantially higher than the other three systems, but it would have retained its most costly rank even if the lower value had been used.

Second was natural gas, with not as much variation as in the coal case. Values were approximately half those of coal. Using the higher value of cost per ton of CO₂ approximately doubled the values of natural gas.

Bioenergy was third, with average values approximately half those of natural gas. At approximately 33 mECU/kWh, German bioenergy was substantially higher than that of the other six nations in the survey. Lowest was that of The Netherlands, at approximately 5 mECU/kWh.

Nuclear was the lowest of the four. Most values were approximately 2 mECU/kWh. France seemed to

have the highest value. The effect of valuing CO₂ at different costs played little role with regard to the relative external costs of bioenergy and nuclear, although all energy systems generate CO₂ either directly or indirectly.

16. CONCLUSION

Many conclusions can be drawn from the hundreds of studies on most aspects of energy risk. Perhaps the most remarkable is the divergence between public perception of energy risk, often centered on the hazards of nuclear power technology, and its often low relative rank in terms of comparative analysis.

Energy risk analysis is highly complex. The subjects discussed in this article represent only part of the field. Energy risk analysis is also somewhat different from other subjects in the field of energy in that much of it is highly contentious, with public debates over assumptions and results. These debates indicate the importance of the field to both policy-makers and the public.

SEE ALSO THE FOLLOWING ARTICLES

Climate Change and Public Health: Emerging Infectious Diseases • Corporate Environmental Strategy • Ecological Risk Assessment Applied to Energy Development • Nuclear Power: Risk Analysis • Occupational Health Risks in Crude Oil and Natural Gas Extraction • Occupational Health Risks in Nuclear Power • Oil and Natural Gas Exploration • Radiation, Risks and Health Impacts of

Further Reading

- American Chemical Society/Resources for the Future (1998). "Understanding Risk Analysis." American Chemical Society/Resources for the Future, Washington, DC.
- Cohen, B. (1990). "The Nuclear Energy Option: An Alternative for the 90s." Plenum, New York.
- Douglas, M. (1985). "Risk Acceptability According to the Social Sciences." Russell Sage Foundation, New York.
- Dyer, J. S., and Sarin, R. K. (1981). "Relative Risk Aversion." Univ. of Texas Press, Austin, TX.
- European Commission (1998). "Externalities of Fuel Cycles: Methodology," 2nd ed., DG XII, Science Research and Development. European Commission, Brussels.
- European Science and Technology Observatory (1999). "On Science and Precaution in the Management of Technological Risk." European Science and Technology Observatory, Seville, Spain.

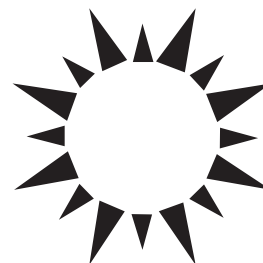
- Franklin, J. (ed.). (1998). "The Politics of Risk Society." Polity, Cambridge, MA.
- Hammond, B., and Coppock, R. (eds.). (1990). "Valuing Health Risks, Costs, and Benefits for Environmental Decision Making." National Academy Press, Washington, DC.
- Hauptmanns, U., and Werner, W. (1991). "Technische Risiken ("Engineering Risks: Evaluation and Valuation")." Springer-Verlag, Berlin.
- Heyman, B. (ed.). (1998). "Risk, Health and Health Care: A Qualitative Approach." Arnold, London.
- Inhaber, H. (1982). "Energy Risk Assessment." Gordon & Breach, New York.
- Leviton, L. C., Needleman, C. E., and Shapiro, M. A. (1998). "Confronting Public Health Risks: A Decision Maker's Guide." Sage, Thousand Oaks, CA.
- Louvar, J. F., and Louvar, B. D. (1998). "Health and Environmental Risk Analysis: Fundamentals with Applications." Prentice Hall, Upper Saddle River, NJ.
- Mossman, K. L., *et al.* (1996). "Radiation Risk in Perspective." Health Physics Society, McLean, VA.
- Paustenbach, D. J. (1995). The practice of health risk assessment in the united states (1975–1995): How the U.S. and other countries can benefit from that experience. *Hum. Ecol. Risk Assessment* 1, 29–79.
- Pochin, E. E. (1983). "Nuclear Radiation: Risks and Benefits." Clarendon, Oxford, UK.
- Rabl, A., and Dreicer, M. (1999). An eye to the future. Long-term health and environmental impacts of energy systems. *Int. Atomic Energy Agency Bull. (Austria)* 41, 31–34.
- Rogers, J. T., and Bates, D. V. (eds.). (1982). "Symposium on the Assessment and Perception of Risk to Human Health in Canada." Ontario Science Centre, Toronto, Ontario, Canada.
- Rück, B. (ed.). (1993). "Risiko ist ein Konstrukt ("Risk Is a Construct: Perceptions of Risk Perception")." Knesebeck, München.
- Sagoff, M. (1985). "Risk–Benefit Analysis in Decisions Concerning Public Safety and Health." Kendall/Hunt, Dubuque, IA.
- Schwing, R., and Albers, W. A., Jr. (eds.). (1980). "Societal Risk Assessment: How Safe Is Safe Enough?" Plenum, New York.
- Sharma, H. D. (ed.). (1990). "Energy Alternatives: Benefits and Risks." Univ. of Waterloo Press, Waterloo, Ontario, Canada.
- Starfelt, N. (1998). How ExternE results can be used for environmental impact comparisons: Sweden as an example. *Nuclear Europe Worldscan* (11–12), 57–58.
- Starr, C. (1969). Social benefit versus technological risk. *Science* 165, 1232–1238.
- Stirling, A. (1999). "On Science and Precaution." Science Policy Research Unit, University of Sussex, UK.
- Uranium Information Centre (2000). "Nuclear Electricity," 6th ed. Uranium Information Centre, Melbourne, Australia. [e-mail: uic@mpx.com.au]
- U.S. Congress, House, Committee on Science, Space, and Technology (1994). "How Safe Is Safe Enough? Risk Assessment and the Regulatory Process, July 27, 1993." U.S. Government Printing Office, Washington, DC.
- Warner, F. (1981). "Assessment and Perception of Risk: A Royal Society Discussion." Royal Society, London.
- Whipple, C. (ed.). (1987). "De Minimis Risk." Plenum, New York.
- Wildavsky, A. B. (1988). "Searching for Safety." Transaction Books, New Brunswick, NJ.
- Wildavsky, A. (1995). "But Is It True? A Citizen's Guide to Environmental Health and Safety Issues." Harvard Univ. Press, Cambridge, MA.

Rocket Engines

WILLIAM ANDERSON

Purdue University

West Lafayette, Indiana, United States



-
1. Introduction to Rocket Propulsion Systems
 2. Fundamental Principles of Chemical Rocket Propulsion
 3. Rocket Propellants
 4. Rocket Engine Cycles

Glossary

bipropellant Propellant system consisting of both an oxidizer and a fuel.

characteristic exhaust velocity A reference term relating achievable chamber pressure to the propellant type and nozzle throat diameter. It is a figure of merit for the propellant and efficiency of combustion.

impulse The total force imparted to a vehicle over the course of the burn time, with units of N-s. The impulse per unit mass is the specific impulse.

nozzle Thrust chamber component comprising converging and diverging sections. The minimal nozzle area is the throat, at which point the speed of the combustion gas is approximately equal to the speed of sound in the gas. The ratio of the nozzle exit area to the throat area is the expansion ratio.

propellant The working fluid in a rocket. The chemical energy contained in the propellant is converted first to thermal energy during the combustion process and then to kinetic energy during expansion in the nozzle.

thrust The reaction force imparted to a rocket primarily due to the change in momentum of propellants as they expand in the thrust chamber.

thrust coefficient The ratio between the thrust and the product of chamber pressure and throat area. It is a figure of merit for the nozzle design.

velocity increment The change in velocity of a vehicle due to its acceleration by thrust (also called ΔV).

Rockets are devices that contain all the elements necessary for propulsion within themselves. They are most useful for space travel and when high thrust rapid acceleration is required. Applications include boosting payloads to low-earth orbit, missiles, satellite station-keeping and orbit transfers, and

interplanetary missions. Rockets are typified by the high velocity of the gas that is accelerated through a supersonic nozzle to generate thrust. They can be further classified according to the propellant state, the thrust level, and the type of engine cycle that is used.

1. INTRODUCTION TO ROCKET PROPULSION SYSTEMS

1.1 Rocket Propulsion

Propulsion is the means by which a force is imparted to change the velocity of a vehicle, affecting both its speed and its direction. In aerospace vehicles, this force is typically called thrust. The thrust applied to a rocket is due to a reaction from the momentum of a high-speed jet directed opposite the direction in which the vehicle is being accelerated. The thrust is proportional to the product of the mass flow rate, $\dot{m}$, and velocity, V , of the jet (i.e., $F = \dot{m}V$). These propulsive forces are used to overcome gravity and drag and to steer the vehicle.

Both rocket and jet engines use jet propulsion to develop thrust. Rocket propulsion differs from its air-breathing cousin mainly in that rockets have self-contained propellants. This feature allows operation at high altitudes and in space. Rockets are typically classified according to the energy source that is used to accelerate the working fluid. Energy sources include pressurized gas, chemical, electric, and nuclear. More exotic types, such as nuclear, solar sails, or beamed-energy approaches, may be feasible in the future, but their consideration is beyond the scope of this article.

Cold gas propulsion comprises a controlled, pressurized gas source and a nozzle and represents the simplest form of rocket engine. Cold gas has many applications in which simplicity is more important than high performance, for example, the

Manned Maneuvering Unit used by astronauts. Nuclear and electric propulsion can be separated into two general categories: thermal and electrostatic/magnetic. Thermal systems work by a transfer of heat generated by the nuclear or electric energy source to a working fluid. Electrostatic/magnetic systems work by accelerating ions through an electrostatic or combined electromagnetic potential. Although some development has occurred, there are no operational nuclear rocket systems. There are operational systems that use electrothermal and electrostatic propulsion.

1.2 Chemical Rockets

The main focus of this article is chemical rockets, whereby the energy that is stored in the chemical bonds of the propellants is first converted to thermal energy through the process of combustion and then converted to kinetic energy by expanding the high-temperature combustion products through a supersonic nozzle. Since the attainable thrust is proportional to the exhaust velocity, and the exhaust velocity is proportional to the square root of the kinetic energy of the exhaust gas, it is readily seen that the chemical energy content of the propellants that can be converted to thermal energy provides a fundamental limit on the thrust that can be achieved.

Chemical rocket propulsion can be categorized according to the original state of the propellant. Rockets that use solid propellants are called solid rocket motors (SRMs), and those that use liquid propellants are called liquid rocket engines (LREs). A hybrid rocket uses liquid oxidizer and solid fuel. LREs and SRMs are the focus of this article.

A general schematic of a pump-driven LRE is shown in Fig. 1. A single gas generator (GG) is used to provide hot gas to turbines that drive the fuel and oxidizer pumps. The pressurized propellants are sent to the GG and the thrust chamber assembly (TCA), which comprises an injector, the main combustion chamber, and the supersonic nozzle. The effluent from the GG is shown to be expanded through separate nozzles; in some configurations, it is injected into the TCA nozzle at an axial location where the gas is the same pressure as that of the GG effluent. Figure 1 depicts a specific LRE cycle—the GG cycle. Other types of cycles are discussed later.

The SRM is one of the most common types of rocket propulsion systems and has a wide variety of applications. An SRM is conceptually a very simple device with very few moving parts. The entire propellant charge, the propellant grain, is contained

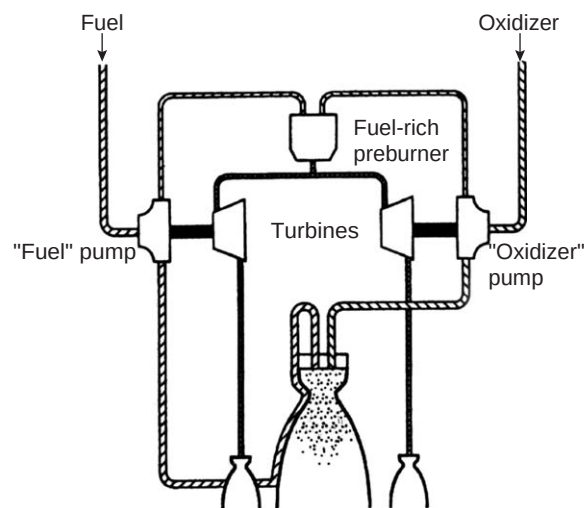


FIGURE 1 Liquid rocket engine schematic. The pump-fed engine shown uses a fuel-rich gas generator to produce turbine drive gases. The turbines provide power to pump the oxidizer and the fuel to high pressures. The high-pressure propellants enter the combustion chamber through the injector and react to form a high-temperature gas. The high-temperature, high-pressure gas is expanded through a supersonic nozzle, in which it reaches a very high velocity.

within a pressure vessel generally called the motor case. The case is protected from hot combustion gases by internal insulation. The nozzle serves to convert the thermal energy of the combustion gases to thrust and may be moved (vectored) to provide a directional force for steering. The polar boss provides a means of connecting the nozzle assembly to the motor case, whereas the thrust skirt provides a mechanism to transmit thrust loads to an upper stage or payload.

Other distinguishing features of chemical rocket propulsion include high operating pressures and high rates of propellant consumption. Operating pressures can be up to 50 MPa for LREs, and flow rates may be >4000 kg/s for high-thrust SRBs. A pump on the Space Shuttle main engine (SSME) could drain an average backyard swimming pool in 25 s, and the energy released by the three SSMEs is equivalent to the output of 23 Hoover Dams. The high energy density of the rocket provides a 100-fold increase in thrust over air-breathing engines of similar size.

1.3 Chemical Rocket Requirements

The energy requirements of a rocket are typically stated in terms of velocity increment or ΔV . The flight velocity required to escape from Earth's gravitational

field (the escape velocity, u_{esc}), neglecting the rotation of the earth, frictional drag, and the attraction of other celestial bodies, can be calculated as $u_{esc} = \sqrt{2g_e r_e} = 11.2$ km/s, where g_e is the acceleration due to gravity at the earth's surface ($g_e = 9.81$ m/s²), and r_e is the radius of the earth ($r_e = 6.38 \times 10^3$ km). Launching an object into space at this initial velocity would not be practical because the object could not withstand the severe aerodynamic heating and dynamic pressures that it would experience in the earth's atmosphere.

For a rocket launched from the surface of the earth where the goal is to put an object into an orbit, the overall velocity increment is the difference between the velocity at the surface and the velocity required for a stable orbit, whereby the centrifugal force of the satellite balances the earth's gravitational field. A circular orbital velocity can be calculated by $u_c = \sqrt{\mu/r_e}$, where μ is the gravitational constant (3.99×10^5 km³/s² for Earth), and r_e is the distance between the centers of the two bodies. The altitude above the earth surface corresponding to low Earth orbit (LEO), which serves as a stable orbit for many satellites as well as a staging point for acquisition to higher orbits, is on the order of 180 km or approximately 100 nautical miles. A typical ideal ΔV requirement for placing an object in LEO, neglecting the earth's rotation and also gravitational, steering, and drag losses, is approximately 8 km/s. Adding these real effects increases the ΔV requirement to approximately 9 km/s. Launching a rocket into an eastward trajectory near the Equator provides the maximum initial velocity and reduces the ΔV requirement. The overall flight velocity budget for a typical Space Shuttle LEO mission launching from Cape Kennedy is shown in Table I. Higher orbits require more velocity increment.

TABLE I
Typical Velocity Increment Budget for Space Shuttle

Circular orbit velocity at 110-km altitude	7790 m/s
ΔV gravity loss	1220 m/s
ΔV turning loss	360 m/s
ΔV aerodynamic drag loss	118 m/s
Orbit injection	145 m/s
Deorbit maneuver for reentry and aerobraking	60 m/s
Correction maneuvers	62 m/s
Initial velocity provided by Earth's rotation at 28.5° latitude	−408 m/s
Total ΔV required	9347 m/s

Round-trips to the moon and Mars require a ΔV of approximately 18 and 27 km/s, respectively.

Other applications for chemical rockets include missiles and orbit transfer vehicles. Ballistic missiles spend most of their flight time in a ballistic trajectory in which only gravity forces are active. Ballistic flight is defined as a “zero-lift” trajectory in which the missile always flies at a zero angle of attack (i.e., the axis of the missile is aligned with its velocity vector). The ΔV for a ballistic missile is determined by its range requirement. Other missile applications include interceptors (e.g., air-to-air and surface-to-air missiles) that must close in on a moving target with high initial acceleration rates and therefore high thrust-to-weight requirements. Orbit transfer propulsion provides transfers from LEO to higher orbits [e.g., geosynchronous orbits (GEOs)] and inclination changes. The Hohmann transfer ellipse is a well-known maneuver between two circular orbits that provides a minimum velocity increment requirement. A LEO-to-GEO maneuver using the Hohmann transfer requires a ΔV of approximately 4 km/s.

2. FUNDAMENTAL PRINCIPLES OF CHEMICAL ROCKET PROPULSION

2.1 The Thrust Equation

The performance of a chemical rocket is governed by the thermal energy contained within the combustion products and the efficiency of the nozzle in converting this thermal energy into kinetic energy and thrust. Figure 2 shows a simple schematic of a chemical rocket. The thrust, F , is the reactive force applied to the rocket and is the sum of the momentum thrust and the pressure thrust:

$$F = \dot{m}V_e + A_e[P_e - P_\infty],$$

where A_e is nozzle exit area, P_e is the gas pressure at the nozzle exit, and P_∞ is the ambient pressure. The momentum thrust is much larger than the pressure thrust. When the pressure in the chamber is sufficiently higher (at least twice) than the ambient pressure, supersonic flow can be achieved in a properly designed nozzle. In a supersonic flow, the velocity increases and the pressure decreases with increasing nozzle cross-sectional area. This process of expansion is the means through which high V_e is obtained.

Although the thrust equation suggests that additional thrust can be obtained when the exit pressure

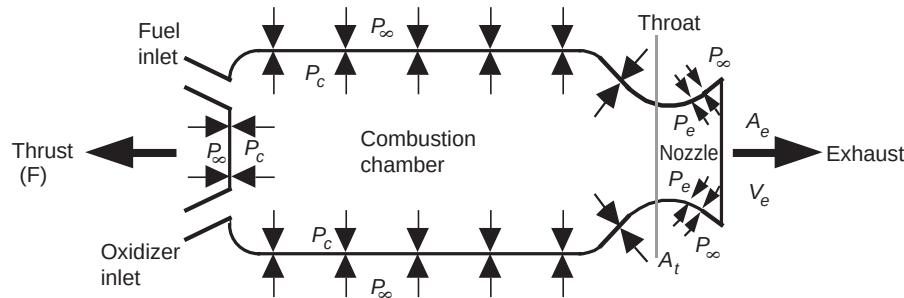


FIGURE 2 Simplified diagram of a chemical rocket. Reaction between the fuel and oxidizer in the combustion chamber produces hot gases, which are then expanded in a supersonic nozzle to high velocity.

TABLE II

Applications and Typical Performance

Propulsion technology	Launch vehicles	Orbit insertion		Orbit maintenance and maneuvering	Attitude control	Typical I_{sp} (s)
		Perigee	Apogee			
Cold gas				X	X	30–70
Solid	X	X	X			280–300
Liquid						
Monopropellant				X	X	220–240
Bipropellant		X	X	X	X	305–325
Dual mode		X	X	X	X	313–330
Hybrid		X		X		250–340
LO ₂ /kerosene	X	X				320–340
H ₂ O ₂ /kerosene	X	X				310–330
LO ₂ /H ₂	X	X				405–460

is greater than the ambient pressure, the loss in the momentum thrust term is greater than the gain in the pressure thrust term. Rocket exhaust nozzles are often designed with an exit pressure equal to the ambient pressure; when P_e is equal to P_∞ , the nozzle is “optimally expanded.” The highest thrust for a given flow rate is achieved when the nozzle flow is expanded to pressures approaching zero in a vacuum environment. Some space engines have nozzle expansion ratios (ϵ , the ratio of nozzle exit area to throat area) exceeding 100. Practically, the expansion ratio is increased until the additional weight of the longer nozzle and the increase in internal nozzle drag losses result in an overall penalty. The physical envelope provided by the launch vehicle also sets limitations on the space engine nozzle expansion ratio.

If the ambient pressure (P_∞) is greater than the pressure at the exit of the nozzle, the pressure thrust term is negative. This loss in thrust is most severe at low altitudes for first-stage engines with exit nozzles

that are configured to produce a maximum amount of thrust over the entire launch profile.

2.2 Rocket Performance Parameters

Mission requirements are often stated in terms of total impulse, I_t , which is a measure of the thrust imparted to the vehicle over the thrusting interval, $I_t = \int F dt$, and has the dimensional units of force-time. Performance is the figure of merit for a specific propellant combination and thrust chamber configuration and is usually related in terms of effective exhaust velocity or specific impulse (Table II). The effective exhaust velocity, c , is defined as $c = V_e + A_e[P_e - P_\infty]/\dot{m}$ and is equivalent to the actual exhaust velocity at optimally expanded conditions. The specific impulse, I_{sp} , with dimensional units of seconds, is the total impulse divided by the amount of propellant that produced that amount of impulse. It is also the ratio of the thrust to the propellant flow rate, $I_{sp} = F/\dot{m}$. The first definition is useful for

SRMs and the second definition is useful for LREs because for SRMs only the total amount of propellant that was burned is typically known, whereas for LREs precise measurements of propellant flow rate at any given time can be made.

Specific impulse is a measure of how efficiently the energy content of the propellants is converted into thrust and can also be defined as the product of the characteristic exhaust velocity, C^* , and the thrust coefficient, C_f : $I_{sp} = C^* C_f$. Definition in this way separates the physical effects occurring in the combustion chamber (C^*) and the exhaust nozzle (C_f). Characteristic exhaust velocity is a parameter primarily dependent on the propellant combination and how effectively the chemical energy contained in the propellant bonds is converted to thermal energy, whereas the thrust coefficient is a measure of how efficiently the thermal energy produced by the combustion reaction is converted to kinetic energy.

Characteristic velocity can be calculated as $C^* = f(\gamma)\sqrt{T/MW}$, where T and MW are the temperature and molecular weight of the combustion products, respectively, and $f(\gamma)$ is a function dependent on the specific heat ratio, γ . The thrust coefficient is an expression that relates the magnification of the momentum thrust due to expansion in the supersonic nozzle, which can be written as $C_f = F/P_c A_t$ and which is dependent only on the specific heat ratio, the ratio of chamber pressure to nozzle exit pressure, P_c/P_e , and the expansion ratio, ϵ , of the nozzle exit area to nozzle throat area, A_e/A_t . These relationships show that high specific impulse is partially attainable by high-combustion gas temperatures and low-combustion gas molecular weights. Further examination of the fundamental thrust coefficient equation would show that higher values of pressure ratio coupled with higher values of ϵ , along with lower values of γ , lead to higher values of C_f .

2.3 The Rocket Equation

For most rocket propulsion applications, the primary requirement is the velocity increment, ΔV , that must be achieved, as discussed previously. The relationship between propulsion system performance parameters and mass fraction is the so-called rocket equation:

$$\begin{aligned}\Delta V &= g I_{sp} \ln \left(\frac{m_0}{m_0 - m_p} \right) = g I_{sp} \ln (m_0/m_f) \\ &= g I_{sp} \ln (1/MR),\end{aligned}$$

where $m_f = m_0 - m_p$ is the final vehicle mass, m_0 is the initial vehicle mass, m_p is the mass of the propellant

consumed, and $MR = m_p/m_0$ is the mass ratio. This equation allows one to calculate the mass of propellant required to reach a certain ΔV based on engine performance and the final mass of the rocket, which is very close to the dry weight of the rocket.

3. ROCKET PROPELLANTS

3.1 Thermodynamics

In a chemical rocket, the propellant serves the dual functions of energy source and the working fluid that provides the momentum thrust. The energy source is the exothermic reaction that occurs between fuel and oxidizer or during the decomposition of a mono-propellant. In an exothermic reaction, the products have lower heats of formation than do the reactants, and the amount of energy that is released, the enthalpy of reaction, ΔH_{rxn} , is proportional to their difference,

$$\Delta H_{rxn} = \sum_{\text{products}, i} n_i \Delta H_i^0 - \sum_{\text{reactants}, j} n_j \Delta H_j^0,$$

where ΔH_i^0 is the molar heat of formation of the i th component, and n_i is the number of moles of the i th component. The heats of formation of elemental molecular species such as oxygen and hydrogen are zero by convention, the heats of formation of typical products such as steam and carbon dioxide are negative, and the heats of formation of unstable molecules such as hydrogen peroxide and hydrazine are positive.

The liberated chemical energy takes its sensible form by heating up the products of combustion. An energy balance between the reactants and products provides the adiabatic flame temperature, T_{ad} ; this is the temperature that the combustion products would reach according to

$$\Delta H_{rxn} = \sum_{\text{products}, i} \int_{T_{ref}}^{T_{ad}} n_i C_{p,i} dT,$$

where T_{ref} is a reference temperature (typically 298.2 K) and equal to the reference temperature for the heat of formation data. It can be seen that the desirable propellant has a high positive heat of formation and reacts with the opposite propellant to form low-molecular-weight and low-heat-capacity products of reaction. Table III summarizes thermodynamic properties of some common propellants and products of combustion.

The equation of conservation of energy may be written for an ideal flow as $h_o = h + V^2/2$, where h is

TABLE III

Chemical Thermodynamic Properties of Common Propellants and Products of Combustion

Substance	Phase	MW (g/mol)	$\Delta H_{\text{form}}^\circ$ (kJ/mol)	C_p (J/mol K)
Al ₂ O ₃	L	102.0	−1620.6	79.0
CO ₂	G	44.0	−394.4	37.1
H ₂ O	G	18.0	−228.6	33.6
N ₂ H ₄	L	32.0	50.6	98.8
N ₂ O ₄	L	92.0	−19.6	142.5
NH ₄ ClO ₄	S	117.5	−295.8	128.1

the specific enthalpy, and h_o is the stagnation enthalpy, which is equal to the sum of the heat of reaction and the sensible enthalpy carried in with the propellants. As the flow is accelerated through the supersonic nozzle, the static temperature decreases. As the temperature approaches zero, h approaches zero. Thus, the stagnation enthalpy sets the limit on the velocity that can be achieved by an infinite expansion through a supersonic nozzle.

3.2 Liquid Propellants

Liquid propellants can be categorized as oxidizers and fuels and as monopropellants and bipropellants. An oxidizer plus fuel is a bipropellant system. Some oxidizers and fuels can act either as part of a bipropellant system or as a monopropellant. Monopropellants generate propulsive energy through the exothermic decomposition of a single molecule or formulated mixture of fuel and oxidizer. Monopropellants allow simple systems but offer low performance. Only three monopropellants have been used in actual flight vehicles: hydrazine, hydrogen peroxide, and propyl nitrate. Monopropellant thrusters are used primarily for attitude control devices and GGs for turbine drive.

The most commonly used monopropellant is hydrazine (N₂H₄). Hydrazine is also a bipropellant fuel hypergolic with nitric acid, nitrogen tetroxide, and hydrogen peroxide. Hydrazine derivatives include unsymmetrical dimethylhydrazine (UDMH) and monomethylhydrazine (MMH), which are more stable than hydrazine with a wider liquid range but slightly lower specific impulse values. MMH is extensively used as a fuel in small attitude control engines. Hydrogen peroxide can also be used as a bipropellant. As a monopropellant, it can be decomposed to form hot water and oxygen.

A bipropellant engine introduces the complexity of a propellant feed system but allows higher performance. Oxygen and fluorine are the strongest oxidizers but must be stored in a cryogenic state. Oxygen is the most commonly used oxidizer in liquid bipropellant rocket propulsion systems. It has a normal boiling point of 90 K with a density of 1.32 g/cm³. Its critical temperature is 155 K and critical pressure is 5.0 MPa. Nitrogen tetroxide is a common storable oxidizer that is hypergolic with many fuels. The Occupational Safety and Health Administration lists a ceiling exposure limit of 5 parts per million with possible health hazards primarily to the lungs and irritation to the eyes, nose, and throat.

The most commonly used storable liquid fuels are RP-1 and hydrazine and its derivatives. RP-1 is a petroleum derivative with a H/C ratio of 1.953, a specific gravity of 0.8, and a MW that ranges from 165 to 195 g/mol. It is typically used with liquid oxygen (LOX) in launch stages because of its combination of high density, low toxicity, and reasonable performance.

Hydrogen is the most common cryogenic fuel. It provides high performance and is an excellent regenerative coolant. However, at its normal boiling point of 20 K, its specific gravity is approximately 0.07; thus, it requires heavy insulation and large tanks. It is especially favored for high-performance, expander cycle upper-stage engines.

In addition to performance characteristics (Table IV), a number of other factors enter into the selection of a propellant for a specific propulsion application. Most of these considerations are highly mission dependent and include propellant density, handling, stability, and transport properties.

3.3 Solid Propellants

The propellant in an SRM must burn at the proper rate to maintain the desired thrust profile and must

TABLE IV

Characteristics of Liquid Propellant Combinations^a

Propellant	O/F	Bulk density (kg/m ³)	C^* (m/s)	T (K)	I_{sp} (s)
LO ₂ /LH ₂	4.83	320	2385	3250	455.3
LO ₂ /RP-1	2.77	1030	1780	3700	358.2
N ₂ O ₄ /MMH	2.37	1200	1720	4100	341.5
N ₂ O ₄ /N ₂ H ₄	1.42	1220	1770	3260	343.8

^a $P_c = 6.9$ MPa; $\epsilon = 40$; $P_\infty = 0$.

have structural continuity to withstand the pressure and acceleration loads introduced during flight. Most physical characteristics cannot be predicted analytically, so heavy use of empirical data is required. Solid propellants are classified into two general categories. In double-base propellants, fuel and oxidizer are mixed on a molecular level. These propellants generally contain a nitrocellulose type of gun powder dissolved in nitroglycerine with additional amounts of minor additives. Both these primary ingredients are explosives containing fuel and oxidizer within their molecular structure. Double-base propellants have been used most frequently in military applications (primarily in ballistic missiles), but their use has declined due to an increased emphasis on safety.

The other general type of solid propellant is called composite propellant to highlight the fact that it is a heterogeneous mixture of fuel, oxidizer, and binder. Fuels for composite propellants are generally metallic powders, with aluminum being the most common. Oxidizers are usually crystalline materials; ammonium perchlorate (NH_4ClO_4), also known as AP, is the most popular. The binder is a rubber-based material that serves to hold the powder/crystal mixture together in a cohesive grain. Current binders include hydroxyl-terminated polybutadiene (HTPB) and polybutadiene nitrile (PBAN). Occasionally, small amounts of very energetic ingredients such as “Her Majesty’s Explosive” (HMX) are added to improve performance.

Solid propellant burning rates are a function of chamber pressure. Most designers assume a burning rate dependence according to St. Robert’s law, $r_b = ap_c^n$, where r_b is the burn rate (cm/s), a is the burn rate coefficient, and n is the burn rate exponent. The constants a and n are determined empirically. Table V summarizes characteristics of some propellant formulations.

TABLE V
Characteristics of Composite Propellants Used in Solid Rocket Motors

Combination ^a	Density (kg/m ³)	C* (m/s)	T (K)	P _c range (MPa)	<i>a</i>	<i>n</i>
A	1840	1541	3636	2.75–7.0	0.415	0.31
B	1800	1527	3396	2.75–5.0	0.399	0.30
C	1760	1568	3392	2.75–7.0	0.561	0.35

^a A: TP-H-1202, 21% Al, 57% AP, 12% HMX, 10% HTPB; Star 63D; B: TP-H-1202, 18% Al, 71% AP, 11% HTPB—Star 48, Star 37; C: TP-H-1202, 16% Al, 70% AP, 14% PBAN—RSRM, Titan IV.

4. ROCKET ENGINE CYCLES

The liquid propellant must be transported to the LRE TCA by some means of force. The more simple method relies on gas pressure to force the rocket propellant through one or more main propellant tank feed valves or directly into the TCA (as in the case of solid rocket motors). The other method is to use engine-mounted hot gas-powered turbopumps to feed the propellant from its tank into the TCA. All large liquid booster engines for all launch vehicles, worldwide, that have ever flown or are flying today are turbopump driven, whereas pressure-fed systems are typically used for smaller propulsion systems, such as spacecraft attitude and velocity control propulsion systems and smaller upper stages.

Pressure-fed systems can be either blowdown or regulated. In the simple blowdown configuration, the propellant tanks are initially pressurized with a finite amount of gas, and the tank pressure decreases as the propellant is consumed. In regulated systems, a constant tank pressure is maintained by controlling the pressurant flow from separate high-pressure gas tanks.

Pressure-fed systems are inherently closed cycles but are limited in performance because of the achievable chamber pressures and hence the velocity to which nozzle flow can be accelerated. The pressurant supply typically ranges from approximately 1 to 4 MPa. The low-pressure range corresponds roughly to the low-pressure limitation of minimum-gauge wall thickness tanks. The high-pressure limit corresponds roughly to the point at which the weight penalty of high-pressure gas supply tanks overcomes the added complexity and weight of a pump-driven system.

Engine power cycles for pump-fed systems include the GG cycle, the expander cycle, and the staged combustion cycle. Generally, the pump-fed cycles require that the work required by the pump be provided by the turbine. The required pump work is related to the flow rate, $\dot{m}_p$, the liquid propellant density, ρ , the required pressure rise, Δp , and the pump efficiency, η_p :

$$W_p = \frac{\dot{m}_p \Delta p}{\rho \eta_p}.$$

The delivered turbine work is related to the turbine flow rate, $\dot{m}_t$, the turbine pressure ratio, TPR, the turbine inlet temperature, T_{ti} , turbine drive gas properties R and γ , and the turbine

efficiency, η_t :

$$W_t = \dot{m}_t \left[\frac{\gamma}{\gamma - 1} \right] RT \left[1 - \left(\frac{1}{TPR} \right)^{\frac{\gamma-1}{\gamma}} \right] \eta_t.$$

Balancing the work flows with the propellant flows can be quite challenging, particularly during engine transients, and requires a significant amount of preliminary analysis and engine system testing.

The GG cycle (Fig. 1) is the most common type of rocket engine used for main propulsion. It is classified as an open-cycle engine and in its simplest form it is composed of a set of turbopumps for each propellant, a GG to provide power to drive the turbopumps, a set of control valves, and the main chamber and nozzle. The gas from the GG passes through the turbines that drive the turbopumps and is then expelled separate from the main chamber. Typically, 4 or 5% of the total propellant flow goes through the GG, which is usually operated at a fuel-rich mixture according to the limitations on turbine inlet temperature. This overboard turbine flow is a loss to the potential performance of the engine. Sometimes the GG flow is used to provide control capability to the vehicle, and sometimes it can be used as film cooling (also known as dump cooling) on the aft end of the nozzle.

The expander cycle is the simplest of the pump-fed engines and can be configured as either an open-cycle or a closed-cycle engine. In either case, the energy to pump the propellants is obtained not from burning propellants in a gas generator but rather from a transfer of heat from the main chamber. Fuel, typically hydrogen, is used as a regenerative coolant before it is used as the turbine drive gas. The thrust chamber and nozzle heat transfer limits the energy available for pumping the propellants and thus the achievable chamber pressure. Furthermore, the fuel must have a high heat capacity and good heat transfer properties and must vaporize easily without passing through a two-phase condition.

The expander cycle is inherently easy to start and control since the dynamics and stability concerns of preburner or GG combustion are avoided. An expander engine starts by using the latent heat capacity of the chamber and nozzle to initially accelerate the turbopumps, prior to main chamber ignition. The expander cycle is ideally suited for upper stages and space transfer applications, in which ease of starting and multiple restarts are required. Operational expander cycle engines used for upper stages include the RL-10-A5 and the LE-5. Thrust levels range from approximately 20 to 40 kN.

The staged combustion (SC) cycle engine (Fig. 3) is a closed-cycle engine that presents the highest performing and most complicated cycle in use. In SC cycles, part or all of the propellants go from the turbopump discharge to a high-pressure preburner, which produces turbine drive gas. The turbine effluent is mixed with the remaining propellants and combusted in the main combustion chamber. All the combustion gases are expanded from a high pressure, so the I_{sp} is high. However, the cycle demands high preburner pressures, which stress the turbomachinery.

The earliest identified SC cycle engine, the Topping Turbine Turborocket, was built by Reaction Motors and produced 5000 pounds of thrust. The turbopump was powered by a turbine located inside the combustion chamber. The drive shaft powering the pumps exited the chamber through the center of the injector. The first SC cycle engine to enter into production was the Soviet NTO/UDMH RD-253 engine, which was used on the first stage of the Proton booster, first launched in 1965. The Soviet Union also utilized the SC cycle on all of the LOX/kerosene NK series engines used for its N1 moon rocket program in the mid-1960s. A well-known SC cycle engine is the SSME, which is probably the most advanced engine in operation today.

The SC cycle can be categorized according to whether the turbine drive gas is either fuel rich or

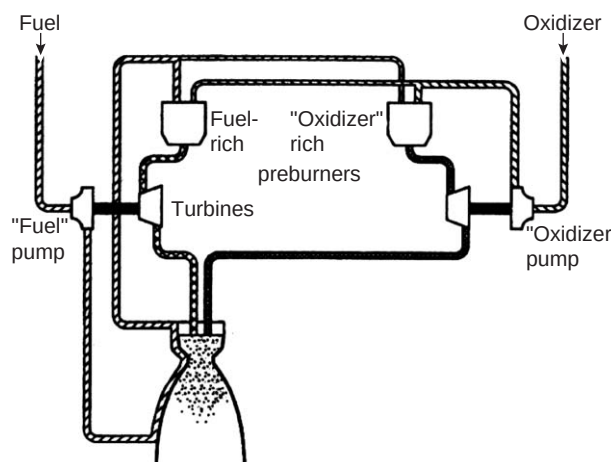


FIGURE 3 Full-flow staged combustion cycle rocket engine. Fuel and oxidizer are pumped to very high pressures before entering the fuel-rich and oxidizer-rich preburners. Prior to entering the preburner, the fuel is used to regeneratively cool the main combustion chamber and nozzle. Warm turbine exhaust gas enters the main combustion chamber, where it reacts further to produce a high-temperature and high-pressure gas that is expanded to very high velocities in a supersonic nozzle.

oxidizer rich. The fuel-rich SC (FRSC) cycle is typically applied to LOX/LH₂ engines (i.e., SSME, RD-0120, and LE-7). The oxidizer-rich SC (ORSC) cycle is used with hydrocarbon fuel-based engines and has been demonstrated in several Russian engines (i.e., RD-170, NK-33, and RD-253) using LOX/kerosene or NTO/UDMH. Another variant of the SC cycle is the full-flow SC (FFSC) cycle, which merges some of the virtues of the FRSC and ORSC cycles but has not been demonstrated in any engine system. The FFSC cycle engine requires two preburners, one producing fuel-rich hot gas to power the fuel turbopump and the other producing oxidizer-rich hot gas to power the oxidizer turbopump.

Although rocket propulsion is a fairly mature field, there is substantial room for improvement. Advancements in the following areas are required: system reliability and long-lived components; engine health diagnostics; high-energy density and storable propellants; lighter, higher temperature materials; nozzles that automatically adjust to

ambient pressure conditions; and advanced thermodynamic cycles.

SEE ALSO THE FOLLOWING ARTICLES

Aircraft and Energy Use • Combustion and Thermochemistry • Internal Combustion (Gasoline and Diesel) Engines

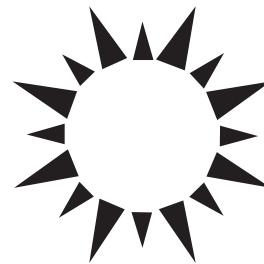
Further Reading

- Hill, P., and Peterson, C. (1992). "Mechanics and Thermodynamics of Propulsion." 2nd ed. Addison-Wesley, Reading, MA.
- Huzel, D. K., and Huang, D. H. (1992). "Design of Liquid Propellant Rocket Engines." American Institute of Aeronautics and Astronautics, Washington, DC.
- Manski, D., Goertz, C., Sassnick, H. D., Hulka, J. R., Goracke, B. D., and Levack, D. J. H. (1998). Cycles for Earth-to-orbit propulsion. *J. Propulsion Power* **14**, 588–604.
- Sutton, G. P., and Biblarz, O. (2001). "Rocket Propulsion Elements." 7th ed. Wiley, New York.

Rural Energy in China

ZHENG LUO

United Nations Development Programme
Pathumthani, Thailand



-
1. Overview of Rural Energy Development
 2. Energy and Rural Environment
 3. Energy and Rural Economy
 4. Energy and Rural Society
 5. Looking Forward

Glossary

chronic obstructive pulmonary disease (COPD) A disease characterized by a progressive airflow limitations caused by an abnormal inflammatory reaction to the chronic inhalation of particles.

global warming potential (GWP) The GWP of one molecule of CO₂ is set = 1, to which the GWP of other gases are placed in relation. For instance, in a period of 20 years, one molecule of CH₄ causes a direct warming effect, which corresponds to that of 35 molecules of CO₂.

ren min bi (RMB) The official currency of the People's Republic of China. One U.S. dollar is about 8.3 RMB.

small hydro In China, capacity less than 25,000 KW is categorized as small hydro.

Substantial changes have taken place in China. Energy development, as both a driving force and a consequence of such tremendous changes, has had profound impact on economic, social, and environmental development. Rural energy has always been a critical issue due to years of energy shortage for both households and industries. Biomass, for long time, has been the only available fuel in many rural areas. A series of policies and strategies have gradually been put in place to encourage alternative fuel exploration and utilization as well as area-based integrated energy systems. As a result, unelectrified rural population dropped from 60% in the 1950s to 5%, while the share of commercial fuel use increased to 43%. The importance of renewable energy has been recognized, not only because of its potential in rural

and remote areas, but also as means of improving the environment and generating economic opportunities.

1. OVERVIEW OF RURAL ENERGY DEVELOPMENT

This article does not pretend to provide a complete in-depth assessment of rural energy in China. However, it is hoped that through a historic summary of the development in past 50 years and analysis of some relevant aspects, this discussion will contribute to a better understanding of the current situation and future challenges.

1.1 Rural Energy Path

Generally, the development of rural energy and rural energy policy in China can be divided into three stages.

1.1.1 Energy Shortage and the Development of Biogas, Small Hydro, and Local Coal Mines (1950–1979)

Despite the high level acknowledgement of the situation, rural energy shortage was severe throughout these three decades. By the end of 1979, the energy consumption per capita in rural areas was 19% of that in urban cities. Seventy percent of rural households lacked cooking fuel, among which 47% had no fuel for more than 3 months per year. Promotion of biogas for cooking and lighting and utilizing local coal resources were strategic decisions from the top since the 1950s; however, not until the late 1970s were specific institutional arrangements and policy measures clarified for a biogas program, with allocation of financial and human resources. Although large-scale biogas projects have been carried out, most digesters failed to perform soon after installation (1 to 3 years of average life time)

due to insufficient technical, managerial, and financial support. Because of this long period of lacking of alternative cooking fuels, regional deforestation, and excessive utilization of agricultural residues, among other factors, have contributed to serious degradation of rural ecological environment, which, in turn, has hampered the rural economic development. The impacts were felt well into the next stage.

1.1.2 Satisfaction of Energy Needs and Development of Policies and Regulations (1980–1990)

After the first national seminar on rural energy in 1980, series of energy policies and regulations have been announced focusing on reforestation, improvement of energy efficiency in both rural industries and households, and the development of renewable energy. The principle of integrated rural energy strategy was defined as developing all possible local alternative resources of energy while rationally utilizing the current sources. Nation wide programs have been carried out in this context, and most of them continuously constitute the current rural energy activities. Thus, it is worth while to have a detailed discussion on the following main results from this period:

Fuelwood Forest. As one of the five forest categories, the main products of the fuelwood forest are wood fuels. Reforestation plans were conceived according to local surveys on fuelwood supply and consumption. From 1981 to 1995, approximately 5 million hectares of forest have been established as fuelwood forest, which, in total, reached 6 million hectares and generated 20 to 25 million tons of wood as fuel per year. However, due to the decline of fuelwood demand, it is expected that the financial viability of these reforestation projects will be highly challenged. Technology optimization and standardization ensuring a diversified and reliable range of products are essential to improve the economic performance. Further more, flexible management and involvement of the private sector are crucial to gradually moving away from subsidies to a market-based operation.

Improved Cook Stoves and Energy Efficiencies in Rural Industries. From 1983, large-scale promotion of improved cooking and heating stoves has been carried out all over the country. By the end of 1999, it was reported that approximately 189 million rural households had been equipped with improved stoves. Generally the efficiency of these stoves, using coal and solid biomass as fuel, has improved by 20 to 40%, which enables energy saving of more than 30

million TCE (ton of coal equivalent) per year. However, to what extend these stoves have been used is not known. Rural industries have experienced a similar process. Improved boilers and kilns, mainly used in tea processing, tobacco drying, and brick and cement producing industries, could have brought about 2 million TCE per year energy saving, as well as significant quality improvement and control in production.

Rural Coal Mining. By the end of the 1970s, coal production from the mines developed at town and county level had already taken a significant role in the energy supply in China. The production of coal from these mines shared 18% of the national total in 1980. The continuous encouragement in collective and private coal mining has tripled their production to nearly 390 million tons and shared 36% of the coal production. Such increase climaxed in the year 1995–1996 when the production shared 45% of national total coal output. Despite low-efficient equipment and technologies and a poor safety record, the rural coal mines have contributed to the rural economy in three ways: easing the energy shortage, ensuring rural productivity, and improving rural employment. While majority of the coal produced from these mines has served market demand locally and from neighboring provinces, a small portion of the product, usually of better quality, has been transferred to the eastern coastal region. To a certain extent, replacing fuelwood with coal has also contributed to forest conservation. The low production and combustion efficiency of coal, however, has had adverse impact on rural environment and health. These effects have also been aggravated by the poor quality of coal in China and the fact that only a small amount has been processed. A case study on Xuan Wei County, Yunnan Province, showed that indoor air pollutants caused by coal combustion was the main health risk factor responsible for the high rate of lung cancer cases in the county.

Rural Electrification. Development of small hydropower has been the dominant strategy for rural electrification in China. In the late 1970s, several policy measures, including subsidy schemes, were adopted to promote small hydro, which resulted in a large increase in generation capacity from an average of 58,000 KW per year in the 1960s to more than 1 million KW in 1979. In the 1980s, local grids and local electricity companies have emerged, which has helped to overcome the disadvantages of standalone systems and also enhanced the rational utilization and distribution of electricity at county level. The

improvement of grid connection continues in the 1990s while the expansion moves toward cross-county connection. Establishment of hydroelectricity enterprises are further encouraged by introducing, among others, a private stockholder scheme. Per capita rural electricity consumption increased 7 times in 40 years and by the end of the year 2000, 28 million people (around 5 million households) in rural areas had no access to electricity, comparing to 450 million in 1978 (nearly half of the total population and 57% of rural population at that time).

1.1.3 Sustainable Energy Development and Market-Oriented Policy and Strategies (1990–2000)

Since the beginning of the 1990s, environmental policy has become one of the principle national development policies in China. As in many other developing and developed countries, renewable energy and energy efficiency have become more prominent in the energy agenda. Several rural energy policies have been published to address renewable energy and sustainable development, energy and environment, and rural energy technology in poverty alleviation. With the deepening rural economic reform since the mid-1980s, energy has been advocated as an integrated component of the rural development and rural market economy. The need for commercialization of conservation technologies, utilization of new and renewable energy resources, and the provision of energy services have been given great emphasis. Progresses in some of these aspects mentioned earlier are detailed in the following discussion:

Renewable Energy (RE). Direct combustion of biomass has gradually decreased both in share and absolute terms, while the utilization of solar, biogas, small and pico hydro, and biomass gasification have increased. Energy policies in addressing environmental problems in the country have set favorable conditions to RE. In particular, several financial measures, such as subsidies, low- or zero-interest loans, taxes, and so on have been essential in renewable energy development. Since 1987, the state council has financially supported more than 500 RE projects in the form of low- or zero-interest loans (more than 400 million RMB, e.g., approximately US\$50 million). From 1990 to 1996, more than 1 million US\$ subsidy was given to training and demonstration of RE technologies, while between 1996 and 2000, subsidies of about 7 million U.S. dollars were given to research and development of RE technologies and implementation. As a result, in

1998, renewable energy utilization, excluding traditional biomass, was 35 million TCE, accounting to more than 2% of the China's total energy consumption (including commercial and noncommercial sources).

Rural Energy Enterprises. Rural energy enterprise emerged with the promotion of biogas in the late 1970s to provide technical services and after build services. Under the strong support from the government on efficiency technologies, various types of devices and equipment have become commercialized. Into the 1990s, different renewable energy technologies have been demonstrated and promoted and, along with it, a more diverse rural energy industry has emerged. By the end of 1998, there were about 4300 rural energy enterprises, 1660 of which were in the industrial sector, manufacturing energy devices for rural users, such as solar water heater, PV, biomass gasifiers, and so on, and more than 2630 were in the service sector, providing technical, marketing, sales, and other services. Compared to 1997, the number of both types of enterprises increased by 8% and 18%, respectively, although the number of total employment has stayed roughly the same (approximately 42,000 to 43,000). Until 2000, such yearly increases have been consistent. However, data show that before 1996, enterprises and employment had been declining in the rural energy sector, and a relatively sharp decrease occurred to the collectively and state-owned enterprises. This short period of decline reflects the further market opening up in rural China, making many inefficient and less flexible state enterprises go bankrupt. Nonetheless, the rural energy industry has yet to mature and favorable policy environment is still essential, since its benefits to overall rural sustainable development have not been integrated and valued fairly (see also Section 3).

The development of rural energy, particularly in the past 20 years, has brought significant changes in the sector in terms of policy, structure, management, technology, and economy. Rural energy consumption increases at a steady rate of 4% per year from 330 million TCE in 1980 to 670 million TCE in 1998. Several comparisons of changes are listed in Table I, while the changes of rural energy structure are presented in Fig. 1.

The fall of rural energy consumption in 1996 is mainly due to the large decrease in the use of noncommercial energy sources (e.g., fuelwood and straw), while the use of coal and oil increased with the latter nearly doubled comparing to 1995.

TABLE 1
1980–1998 Rural Energy Development in China

	1980	1998
Rural population (million)	796	869
Average rural income per person (RMB)	191	2162
Total energy consumption (million TCE)	330	670
Average energy consumption per capita (Kg CE)	420	770
Total electricity, consumption, including household and rural industry (billion KWh)	13	129
Average household electricity consumption per capita (KWh)	8	68
Population without electricity (million)	450	60
Clean fuel for cooking (electricity, LPG, coal gas, biogas, producer gas, solar, etc.) (million households)	3.7	56.3

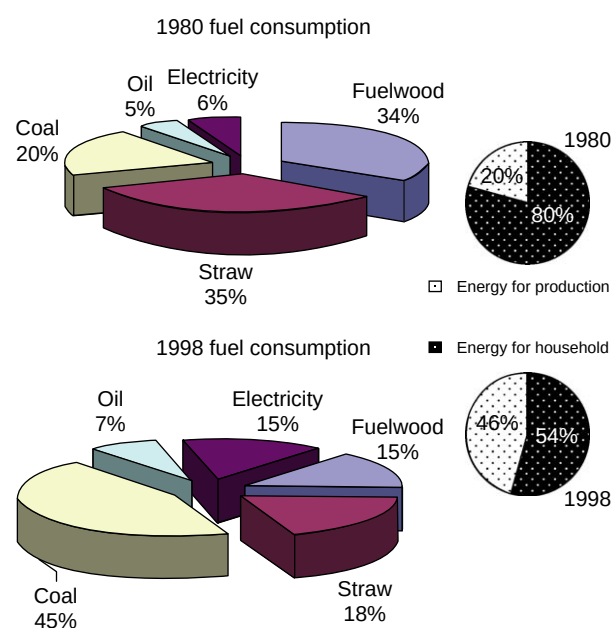


FIGURE 1 Fuel consumption structure changes in rural China 1980–1998.

Decrease in residential electricity consumed in 1996 was uncertain. Official figures from the Ministry of Agriculture show that the consumption was 29 million TCE, which was two-thirds of that in 1995 and, although increasing in the following years, it has not regained the level of 1995. However, according to the statistics from Department of Rural Electricity of the National Electricity Power Bureau, no such decrease has been observed. Nevertheless, energy

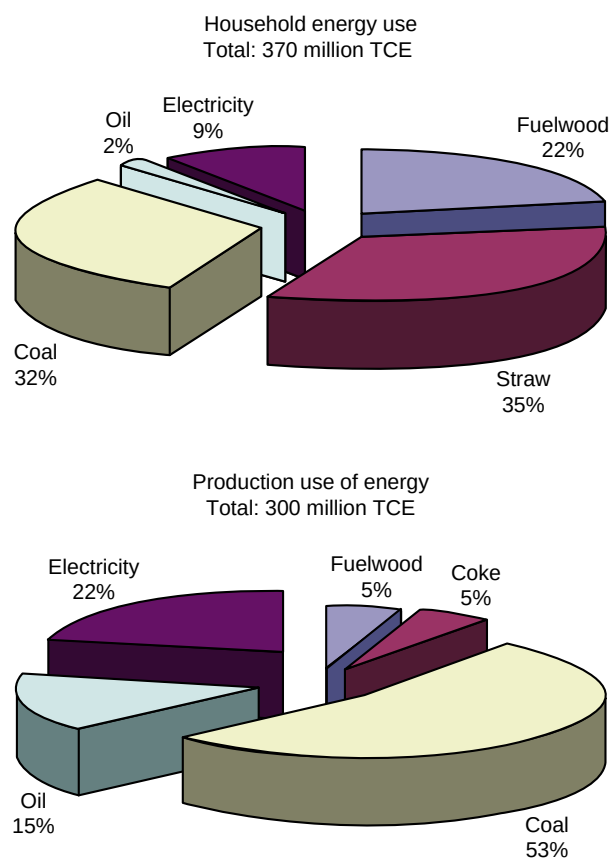


FIGURE 2 Rural energy consumption, 2000.

consumption in rural households has been grown in slow pace since 1995, if not decreased, mainly due to the closure of small coal mines and power producers. In the same period, rural productivity use of energy grew about 16% despite the successful efficiency program in rural industries nationwide. This has significantly contributed to higher income and improved livelihood in many rural areas in China.

1.2 Rural Energy Today

Since 2000, rural energy has accounted to approximately 45% of the total energy consumption and 35% of total commercial energy consumed (e.g., excluding fuelwood and straw). The total rural energy consumption was 670 million TCE in 2000 (see Fig. 2). The majority of rural households still use solid fuel for cooking and heating, while 84% of urban households are using gaseous fuels (LPG, natural gas, and coal gas). China has undergone series of reforms in the energy sector in the past 10 years, and the reform is being extended into rural power grid. The long-term goal of the power sector is

the establishment of a unified national grid. It would require a modern power market in which small and large plants are able to sell electricity to the grid at market-determined price. It would also require a well-defined regulatory role of the government. In the short to medium term, many rural areas will stay with local grid or isolated (village) grid, where renewable energy sources are most promising.

1.2.1 Energy Efficiency

As already mentioned, energy efficiency has been improved tremendously in the past years. However, comparing to national and international averages, rural industries still consume more energy per product unit as shown in Table II, and a significant difference exists between the western and the eastern regions, where efficiency is in general much higher. Because the out-of-date technologies and machinery still prevail in agricultural activities, energy consumption for mechanical power is very high, accounting for nearly half of the diesel used in the country. The current energy efficiency program, focusing on the modification of old equipment and technologies, results in diesel saving of 300,000 tons

TABLE II
Energy Efficiency of Selected Rural Industries

Type of product (energy consumption/ unit)	Average of rural industries	National average	International average
Brick (kgce/10,000 pieces)	1093	780	780
Cement (kgce/t)	161	133	94
Coke (tce/t)	1.56	1.1	1.1
Pulp and paper (tce/t)	1.55 ~ 1.7	1.25	0.51
Electricity (from coal, gce/kWh)	450 ~ 550 (< 50 MW)	320 ~ 350 (300 MW)	329 (Japan)

per year. However, the demand and consumption of diesel in agricultural sector will continuously increase since mechanical power is only responsible for approximately 50% of the productivity.

1.2.2 Clean Energy

With the stabilized balance between demand and supply of energy, China has been moving forward to clean and modern energy and technologies. The demand for clean fuel (electricity, LPG, coal gas, and modern renewable fuels) from rural households has five-fold in the past 10 years, and this trend will continue. Currently, renewable energy and modern renewable energy (including small hydroelectricity) account for 19.8% and 2.3%, respectively, in the total national consumption, as shown in Fig. 3.

A detailed breakdown of the use of renewable energy resources in 1998 is presented in Table III, which shows that all new renewable energy (excluding traditional biomass and all hydroelectricity) constitutes 0.3% of total national energy consumption (i.e., 5.4 Mtce). According to national planning on renewable energy development, this figure is expected to reach 2% by the year 2015, which indicates that total consumption of new renewable energy would have to reach approximately 60 Mtce in 15 years, which may require hundreds of millions of U.S. dollars, as subsidies from central government. Thus, the commercialization of renewable energy and technologies, with a well-performing rural energy service sector, would be crucial. The existing rural energy industry covers almost all renewable energy technologies; however, their stage of development and commercialization varies greatly and are presented in detail in Section 3.

Although domestic coal meeting the demand will still be the strategy in rural energy development in the foreseeable future, coal combustion, particularly at the small scale, has been and will be unavoidably having a negative impact on rural environment and health. Modern renewable fuels, such as solar water

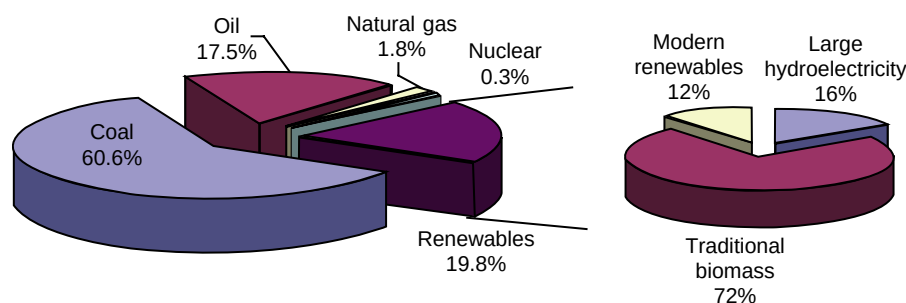


FIGURE 3 Renewable energy consumption in China.

TABLE III
Utilization of Renewable Energy Resources in China (1998)

Resources	Utilization	Mtce
Hydro electricity — small hydro ^a	208 billion kWh, 80 billion kWh	77.6, 29.84
Biomass	221.98 Mtce	221.98
Biogas	2.36 billion M ³	1.73
Bagasse for power	800 MW/2000 GWh	0.75
Geothermal for heat	0.4 Mtce	0.4
Geothermal for power	27.8 MW/110 GWh	0.05
Solar water heater	15 million M ²	1.8
Solar homes	15 million M ²	0.37
Solar cooker	244,000	0.04
Solar PV	13 MW/28.7 GWh	0.01
Wind	242 MW/641 GWh	0.24
Total		305

^aNote: Micro hydro not included.

heaters, biogas, producer gas from agricultural residues, and electricity from wind and microhydro, could provide some viable alternatives for rural and remote areas. They have the potential in competing with the high cost of LPG and coal gas and meanwhile improving indoor air quality (see also Section 2).

A number of new and renewable energy projects, under the Tenth Five-Year Plan, are currently being carried out at nation wide. Since these projects are and will be of significance to the long-term development of the rural energy sector, it would deserve some discussions.

1.2.2.1 The “Brightness” Project By the end of the last century, about 7 million households, accounting for 3.55% of total rural households, have yet to be electrified. Most of these households are located in the western region and the majority of them are of minority origins. To address this situation, the “Light” project was initiated by the national Planning Committee and supported by several other countries [e.g., Japan, Germany, the Netherlands, the United States, and the United Nations Development Programme/Global Environment Facility (UNDP/GEF).] The 10-year target of the project is to provide electricity, in average 100W/person, to 23 million people in remote areas with renewable technologies such as wind and solar PV. The first period of the project (from 2000 to 2005) will install 1.78 million sets of household systems

and 2000 sets of village systems, which will require, approximately, a total of US\$1.2 billion. It is expected that the majority of the investment will come from the users themselves, while only 6.7% will be provided by the central and local government by means of subsidies, especially for the village systems and service networks. The project, in its first period, will focus on the development of domestic industries of wind generators and solar PV; the establishment of sales and a after sale service network; technical training on installation, optimization, repair and maintenance; standardization; and quality control mechanisms.

1.2.2.2 Biomass Gasification Demonstration Project With 65 million rural Chinese without electricity, 70 million are in shortage of cooking fuel. As shown in Fig. 2, agricultural residues, mainly straw, account for 35% of rural household energy use. The average efficiency of straw directly used for cooking is just 5 to 8%; however, after gasification, the overall efficiency reaches about 35 to 45%. Most of all, converting solid fuel to gaseous fuel could bring huge environmental, social, and economic benefits, such as improved indoor air quality and health, reduced of load of house work for women, the sense of modernization of rural communities, and reduced greenhouse gas emissions. However, these externalities are not internalized financially; thus, in average the daily expense on fuel in terms of producer gas is 7 to 8% higher than that in terms of coal gas. One of the first demonstration projects was initiated in Shangdong province with a gasification system and a village gas pipe line installed in 50 villages. It brought significant benefits to those villages in terms of reducing smoke and liberalizing women from the drudgeries of fuel handling. From 1999 to 2000, the national planning committee budgeted 12.5 million RMB (~US\$1.5 million) for wider dissemination of the technology in seven other provinces. It is expected that 53,900 households would be using producer gas for cooking. Standardization and commercialization of gasifiers and gas stoves are key issues to be addressed, along with the demonstration projects (see Fig. 4; see also Section 2).

Modern energy is the main engineer of economic growth. Meanwhile, national policies and measures to promote such growth have a profound impact on sustainable energy development. One of the strategic developments focuses is the current Western Development Plan. Through rational and efficient exploration, distribution, and utilization of resources, the development plan aims at stimulating



FIGURE 4 Biomass gasifiers manufactured and used in Yunnan, China.

economic growth in the West and minimizes the social and economic disparities between the western and eastern regions. Western China is rich with both renewable and conventional energy resources. However, the demand of energy mainly comes from the Eastern coastal areas. Two major programs under the development plan are to transfer both natural gas and electricity (generated from coal) from the western to the eastern region. Mostly energy resources are located in rural areas and the pipe line and network will be most likely constructed along remote villages and rural towns, thus, it is hoped that these programs would enable those rural households to access clean energy services. These programs would also encourage the growth and development of energy service industries in the western region, which could provide employment and income generation opportunities.

The previously discussed rural energy development in China has had a profound impact on the sustainable development in all of its three dimensions (i.e., environment, economy, and society). The impacts are complex and interrelated. The following sections examine several characteristic issues in each of the three dimensions, and through that, the challenges and possibilities in advancing rural energy development are discussed.

2. ENERGY AND RURAL ENVIRONMENT

Small-scale coal mining has caused serious environmental problems in rural China. As mentioned in the previous section, these mines are mostly collectively or privately owned by the counties; little capital was invested in technology and equipment. The average efficiency of coal production (e.g., resource recovery)

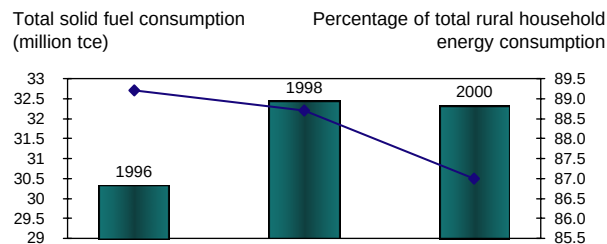


FIGURE 5 Reduced solid fuel consumption in rural household.

is much lower than 50%. Most of the mines are very labor intensive and some have the productivity as low as 1t/labor. Few environmental control measures have been adopted in these mines, and only a small portion of coal produced has been further processed with, for instance, de-sulfur technology. From the end user side, rural households and industries are the main consumers of the low-quality coal from the small mines. Rural industries are known to have outdated equipment with low efficiency (as presented in Table III). This situation has worsened as many high pollutant and energy intensive industries have moved to rural areas. In average, the brick, cement, paper, coking, and casting industries in rural areas share 8% of the total production of the country, while they account for more than 50% of the total energy consumption in these industries. For rural households, since the late 1990s, the consumption of solid fuels (e.g., biomass and coal) has increased only marginally and their share in the total household energy use started decreasing, as shown in Fig. 5. This indicates that with the promotion of improved technologies, efficient stoves, coal, and biomass energy intensity decreased and resulted in energy saving and, consequently, emission reduction both indoors and outdoors. However, it is difficult to assess the actual changes and effects due to the lack of detailed data.

Overexploration of fuelwood still persists in some rural areas despite the relatively successful reforestation program since the 1970s. From the 1990s, fuelwood supply and consumption has been roughly kept in balance due to mainly improved energy efficiency and promotion of coal replacing fuelwood; however, statistics show that fuelwood consumption is higher than the sustainable supply at various degrees in more than ten provinces. For example, Yunnan Province, located in the Southwest of China, has relatively successful solar heating and biogas programs. However, data in 1999 show that fuelwood consumption was about 32 million M^3 while the total sustainable supply from all type of forest was only about 14 million M^3 . Several counties in Yunnan

province are located upstream of, among others, the Yangtze and Mekong River. Years of over exploration of the forest for timber and fuel in these areas have resulted in soil erosion and degradation.

In 1998, agricultural plantation generated approximately 600 million tons of residues, of which 27% was utilized mainly as animal feed and for industries, 15% was returned to the field as natural nutrients, and 47% was used as fuel. On one hand, the declined consumption of agricultural residues as fuel has made the returning of residues back to the field possible. On the other hand, this has led to about 65 million tons of residues left for disposal. Since decomposing stem residues is a rather long process, most of it is burned openly to prepare the field for the next crop. In many areas of central China, this has become a serious seasonal environmental problem.

2.1 Rural Energy and Emission Reduction: The Potential of Biomass Gasification

Analysis has confirmed that China actually has reduced its greenhouse gas emissions, even as its economy has grown rapidly (as shown in Fig. 6). However, due to lack of data from monitoring and evaluation, it is difficult to analyze the share of the emissions from rural energy consumption. Estimates of all energy efficiency projects claim more than 32 million tce energy saving capacity (30 million tce from the improved fuel/stove projects in rural homes) per year, which would be responsible for about 70 million tons CO₂ avoided, that is, two-thirds of the CO₂ reduction from coal in 1998–1999 and twice of the total CO₂ reduced of the year. The steeper CO₂ reduction from coal indicates higher coal efficiency (comparing with coal consumption) and the increases of CO₂ emission from other sources.

As discussed previously, biomass gasification provides a promising alternative in replacing solid

fuels for cooking and heating in rural areas. The amount of agro-residue available for energy use in 1998 was about 350 million tons (possible losses already taken into account), 58% of the total residues generated. Since 285 millions tons of the residues are already used as fuel, it could be assumed that all the available residues could be gathered and transported without imposing any major changes to practices and workload. Assuming 350 million tons of residues could be converted to gashouse fuel for rural cooking and heating, their contribution to emission reduction can be estimated. The average emission factors of coal and biomass fuel for household applications are shown in Table IV.

Based on technical specifications of producer gas in China, it is assumed that gas production rate is 2 M³/kg residues and 1 M³ of producer gas would replace 0.25 kgce; thus, 350 million tons of residues would generate 700 billion M³ of gas and equal to 175 million tce.

When producer gas replaces all the residues used for fuel (285 million tons) and partially replaces fuelwood (i.e., 32 million tce), it could avoid, in total, 109 million tons of CO₂; this is about the total emission reduction from coal in 1998–1999. When producer gas replaces all the residues (285 million tons) and partially replaces coal (32 million tce), it could avoid, in total, 167 million tons of CO₂.

2.2 Rural Energy and Health: A Possible Win-Win Strategy

The simple household solid-fuel stoves commonly used in rural areas do not obtain high combustion efficiency. They emit a substantial amount of fuel carbon as byproduct of incomplete combustion (PIC)—such as carbon monoxide (CO), methane (CH₄), and total non-methane-hydrocarbons (TNMHC)—as well as sulfur dioxide (SO₂) and

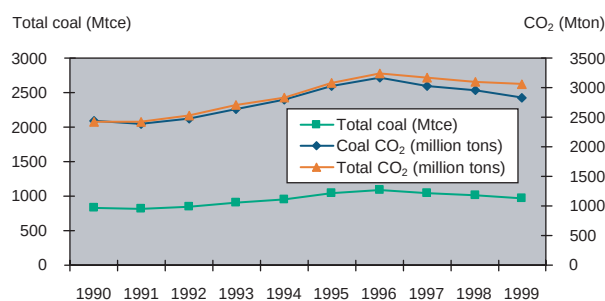


FIGURE 6 China CO₂ trends.

TABLE IV

Average Emission Factors of Household Fuels in China

Fuel	Overall efficiency (%)	g CO ₂ /kg fuel	g CH ₄ /kg fuel	Total g CO ₂ /kg fuel ^a
Crop residues	14	1130	5	311
Wood	23	1520	5	311
Coal	35	2280	3	2466

^aNote that wood and residues are assumed to be CO₂ neutral. And in a 20-year time horizon, the Global Warming Potential (GWP) of one CH₄ molecule is 22.6 times higher than that as CO₂.

particulates. Some of these PICs are hazardous to health when inhaled under the concentrations commonly found in homes using unvented biomass or coal stoves. As discussed, China has carried out, in the past 20 years, efficient cooking stove dissemination programs, which claims that the various technologies have benefited 184 million households by the end of 1998, and this figure is increasing at 2.2% per year. However, again, no monitoring and evaluation data are available to exam the true effect of the stove program, particularly in terms of health improvement. With improved efficiency, carbon emissions could be reduced as discussed in the previous section. It could also result in fuel saving, if more energy is actually delivered in less PICs or if more carbon is converted to CO₂. Table V presents data from site analysis of Chinese stoves. The data show that improved efficiency does not necessarily accomplish emission reduction. However, processed and high-quality fuel is always preferable.

A study of the health benefits from emissions reduction in China has given some further interesting results. It was found that greenhouse gases (GHG) reductions resulting from changes in energy use are generally accompanied by substantial short-term health benefits. However, the degree of health benefits varies greatly with the choice of energy technologies and sectors. The variation appears to be larger when shifting between sectors rather than between technologies, and the largest benefit can be obtained in the household sector. Even conservative estimates show, for example, that the health benefits of one tonne reduction in particulates emission from household stoves are at least 40 times larger than those from coal-power plants. This study also shows that although GHG reduction itself is usually accompanied by increasing costs, when the economic

benefits of improved health are counted, there is a net economic benefit for those household energy options, involving a shift from traditional use of coal and biomass. This strongly suggests that emission reduction in the household sector can be a real win-win strategy for China.

3. ENERGY AND RURAL ECONOMY

As essential to industries, modern energy services are prerequisites for rural economic development. Most rural industries and agro-product processes have been limited to solid fuel (biomass and coal) with outdated equipment as main energy input. This, as already discussed, has resulted in unnecessary waste of energy and constraints to productivity improvement and products diversification. Figure 7 is a simplified diagram showing the inter linkages of lack of energy services and its impact on agriculture and local economy, which is complex and dynamic.

By no means, this section could provide a comprehensive overview of such complexities in rural China. Rural energy production, supply, and services as part of the rural economy have also undergone tremendous changes in terms of market potential and development, and the following will exam this aspect in detail.

3.1 Rural Energy Business

As discussed before, since 1996, the number of rural energy enterprises, as well as their market value and employment, has been increasing consistently. The current rural energy industry covers almost all renewable energy technologies; however, their stage of development and commercialization varies greatly. According to China Rural Energy Industry Association, the technologies and products are grouped in six categories and are under the mandate of six committees in the association. The market situation of these technologies/products is resented with some details.

Solar heating—including Solar Water Heater (SWH), Solar Cooker, Green House and Solar Homes—has been developed rapidly in the past 20 years in China and has become the main product of the sector. There are two main types of SWH on the market (flat-panel and vacuum tube), with an average heat production efficiency of 100~150 kgce/M². The total sale of the two types was 5 million M² in 1999, which was an increase of 25% from 1998. Total installed capacity of SWH has reached

TABLE V
Health Hazardous Emissions — Improved Fuel/Stove

Fuel/stove	Efficiency	g PIC/MJ delivered	g NO _x /MJ delivered	g SO ₂ /MJ delivered
Unprocessed coal	37%	4	0.2	0.02
Coal briquette	14%	24	0.04	0.04
Wood/improved	24%	20	0.1	—
Wood/brick	13%	12	0.2	—

PIC = CO + CH₄ + TNMHC-C (carbon in total none-methane hydrocarbons) + TSP (total suspended particulates).

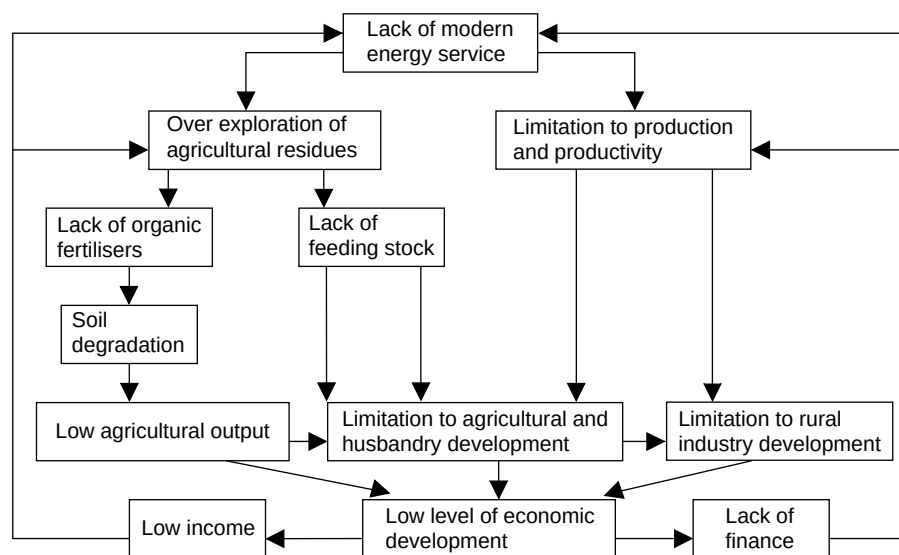


FIGURE 7 Example of energy shortage and its impact on rural economy.

20 million M^2 , of which more than 50% is in rural areas. There are nearly 1000 factories with annual production capacity of 6 million M^2 , which values more than US\$600 million. In total, 20,000 people are employed by these enterprises. By the end of 1998, solar home system (SHS) construction reached 15 million M^2 in total, from which 6.5 million M^2 is located in rural areas. The rate of growth for SHS was 54% per year in the past 10 years. On average, SHS could accomplish 75 to 90% fuel savings on heating. Figure 8 shows a flat-panel solar water heater.

Although there is wide range of products in the energy-efficient cooking devices sector, only two have achieved a large market share. One is the drinking water heater, with a yearly production and sale of more than 300,000 and a value of more than US\$12 million. There are more than 1000 factories, mainly small- and medium-sized enterprises (SMEs). The other one is coal-saving stoves. There are five factories reaching an annual production of more than 10,000 stoves.

There are more than 2600 biogas enterprises all over the country with nearly 18,000 employees. The service provided includes digesters (both large scale and household scale) and its design, necessities, and end use devices, technologies, construction, consultation, repair maintenance, and so on. In total, about 7.6 million digesters have been installed in rural China. Still in technology demonstration stage, there are 740 large and 49,000 medium digesters installed in large farms, industries, and small cities. Although



FIGURE 8 Flat-panel solar water heater, Yunnan province.

small in size, the market for household biogas systems could be promising. Every year 300,000 new digester are being built, and another 100,000 need to be replaced, which implies approximately US\$70 million per year. However, one of the key issues in commercializing small digesters is standardization. Semiconstructed digesters have been developed to secure consistent quality as well as reduce the time needed for construction.

Another technology is solid biomass conversion equipment, including briquetting, direct high efficiency combustion, gasification, liquadization, and carbonization equipment and facilities. Although these products are still in the stage of research and development and demonstration, some 50 factories have already been in production and their products are sold in several provinces. A study of the market potential of a village scale biomass gasification system shows that the average income level of a village

affects the current market of the technology. In the case study, it is shown that when yearly household income reaches 10,000 RMB (about US\$1200), residues as fuel are abandoned and the financial implication of gasification is more acceptable.

A new type of liquid fuel/stove presents yet another technology. The liquid fuel in this category is mainly referring to methanol and ethanol from the waste of oil and chemical industries. There are about 600 research and production enterprises in this sector, among which several are more than \$1 million enterprises. Although the technology is yet to be standardized and further improved, it has already reached more than 300,000 households.

Small-scale electricity production makes up another category. This sector includes small wind, solar, and hydropower industries. More than 50 enterprises are engaged in small wind power research and development and equipment production. Currently, the production of small wind turbines is about 20,000 units per year. Products range from 100W to 5 KW, of which the 100 to 300W types are most in demand. The technology has matured in the past 5 years and production quality has been stable. Solar PV industry has been commercially established. There are more than ten solar PV production lines in China, each with 4.5 MW per year capacity. The annual production has been around 2 MW. In total, 12 MW has been installed, of which 5 MW in rural areas. In average, the system cost is around US\$10 to \$12 per watt at peak and electricity could be generated at 30 U.S. cents/KWh. Other products associated with the systems include lighting appliances, batteries, and pumps. More than 100 enterprises are engaging in their production, mainly located in Gansu and Qinghai provinces. A small hydro industry has been developed to a relatively large scale with approximately 1 million employees. Meanwhile, the mini hydro industry is in expansion with annual sales of 15,000 systems, mainly in the Southwest provinces.

The development of these industries depends largely on the economic viability of these technologies in reality, which is influenced by several factors: the abundance of the energy resources, the demand and willingness to pay by end users, system configuration and performance, system costs, and financial and policy scenarios. Under these factors, the following specific and direct limitations have been identified based on past project experiences in rural China: There are no unified technical standards. Thus, it is difficult to guarantee the product quality with different standards in different regions. In some

cases, markets and consumers are confused with bogus products of low quality; most of the companies are inexperienced in getting loans from banks. Therefore, the shortage of capital is a major barrier for companies to expand their sales and improve after-sales services. With China joining the WTO, increasing domestic and international competition is expected, especially for the PV and wind sectors due to their relatively low level of production technology. Markets need to be closely examined and regulated with proper financial means in order to sustain and develop rural energy enterprises, as well as to reflect their true economic benefits.

3.2 Rural Energy and Income Generation: The Biogas and Eco-Farm Model

As mentioned, biogas technology, for providing cooking and lighting fuel for rural households, was disseminated in the 1950s. However, due to the immaturity of the technology and lack of management and maintenance skills, most of the digesters failed to perform soon after being installed. Similar failures were experienced in the 1970s. More than 7 million units were built in just a few years. Despite the potential environment and health benefits of biogas technology, its economic performance was recognized as a key to success. Two variations of biogas system in combination with eco-farming model have been gradually developed in the past 20 years, the Southern model and the Northern model. The Southern model focuses on the utilization of the sludge from the digester as fertilizer, thus, animal stable, a biogas digester, and fruit/vegetable plantation have become components of one "eco farm." On average, production cost per hectare was reduced by US\$360. Financial payback of the system is about 4 years. Due to low temperature, biogas use in the north had been limited. In the Northern model, animal stable, toilet, biogas digester, and green house are built on one spot to ensure sufficient and better temperature for animal growth and anaerobic reaction. The initial investment of the Northern model is about US\$480 and the payback time is usually within 2 years. According to survey, households installing the Northern model were benefited with US\$380 increased income per year. Currently, 210,000 and 81,300 households have adopted the Northern and Southern models, respectively. They produced more than 4 million tons of fruit and vegetable without using chemicals in 1998. By the end of 1999, 6.7 million households were provided with digesters, and their total gas production is around 2 billion M³ annually.

4. ENERGY AND RURAL SOCIETY

Economic reform is changing the social structure and lifestyle of rural population. As part of the overall development in rural communities, the increasing availability of improved energy services has contributed to improved living and working conditions, better education facilities, saved time for other activities and leisure.

4.1 Rural Energy and Gender Equity

Based on the law of China, women and men are equal. While this has greatly increased the social significance of women in the cities, such development in rural areas has been lagging behind. Rural women have been, for centuries, taking care of household duties. They are not only responsible for the well-being of children and other members of the family, but also for the basic daily livelihood, such as food, cooking, feeding, and washing, in addition to helping in the field. With the broadening of economic reform in rural China, mobility and business opportunities largely increased, which led to more and more agricultural laborers, particularly men, transferring to nonagricultural sectors and working in peri-urban and urban factories. As a result, more responsibilities fell upon women. Thus, unless the household burden is alleviated, there will be little chance for true gender equality in rural society. For many rural women, fuel for cooking and heating has been drudgery in terms of time and labor spend on fuel collecting and smoke irritation from fuel burning. A typical rural family of five would require 3000 kg of firewood per year, which would need 80 days to collect, an average of 5 hours per day. Plus another few hours to cook food and feed, many have to work at least 15 hours a day. Women as users have deeply realized the benefits of improving their fuel situation and have been actively initiating and participating in rural energy projects.

The Forest Research Institute of Yunnan province in Southwest China has carried out a comprehensive research on women and rural energy based on case studies in nine different counties in the province. (In Yunnan province, the Forest Department of the provincial government is in charge of rural energy development.) The findings of the study show that although the form and type of technologies vary from project and region, several common achievements can be identified as follows:

- The energy project has brought great benefits to women in terms of health and living environment.

The physical workload has been reduced, and in many cases less fuel consumption resulted in savings and more time available for increased opportunities for other activities.

- Decision making on fuel and household expenditures in rural China mostly relies on women. Thus, knowledge about energy technologies and related issues is as essential as active participation in energy projects. These energy projects have provided opportunities of training and self-improvement for many women, which in turn have ensured the success of project/technology promotion and implementation.
- Engaging women in energy projects has been seen as one of the means to increase their self-confidence in taking part in community development and getting involved thereafter in all other fields of rural development, some even becoming a leading force in rural economic reform.

4.2 Rural Energy and Community Development

The lifestyle of rural communities links closely to energy production and use. Issues discussed in previous sections all have a profound and complex impact on rural communities. The availability of modern energy services, such as electricity, heated water, and liquid cooking fuel, has affected rural life as it has in urban areas. It has contributed to the modernized agriculture and rural industries, improved the indoor air quality, enriched life with new information and opportunities, increased the speed of daily rhythm, and changed traditional habits and ideas. Closure of rural coal mines unavoidably seriously affected employment and income generation, directly and indirectly, since these coal mines employed more than 3 million rural laborers and were the main source of fuel for most rural industries where more than 20 million were employed. Moreover, coal has been the economic foundation of half of the coal-producing counties. The communities in these regions are in transition. Social security and training for reemployment are main issues of concern, among others.

Meanwhile, evolution of rural society, such as reduced family size and improved living standards, has in turn challenged the energy supply, technology, and policy. Stove programs in the past have helped millions of households to build stoves for large multigeneration families. Many of them are not used regularly, particularly when gaseous fuels became available and affordable. It is often observed that

various types of fuel are used in rural household due to mainly economic and security reasons. More and more agro-residues have been left unused when better quality fuel was available and higher living standard was achieved. The ban of open burning of residues would have to be combined with control and enforcement, which is rather costly. Also it does not provide alternatives to utilize the materials. As already discussed, gasification technology has been demonstrated to be a viable solution. One of the key issues on village scale application is the management of a gasification system. Village communities should be mobilized to either direct management or cooperation with service providers. Fuel supply policies, financial incentives, and energy market should be designed with these issues in mind.

5. LOOKING FORWARD

China's Tenth Five-Year Plan (2000–2005) has set several targets in relation to rural energy. By the year 2005, annual modern renewable energy utilization is expected to reach 13 Mtce (excluding hydroelectricity and traditional biomass). This would avoid 10 million tons of CO₂ and 600,000 tons of SO₂ and particulates emissions. It would also electrify 5 million to 6 million people and provide employment for 200,000 people. Annual production capacity and total installation of solar water heaters are expected to be 11 million M² and 64 million M², respectively, and for PV, 15 MW, and 53 MW, respectively. Wind power installation is expected to reach 1.2 million KW. Annual production of equipment would be equivalent to 150~200 MW capacity. Finally, large-scale biogas application in industries and husbandry and biomass gasification are expected to provide 2 billion M³ fuel gas.

Compared to energy use in 1998, modern renewable energy utilization will more than double. Although consumption of modern renewables has been increasing fast, financial support remains crucial given the current immature development of the market and business environment. Experiences show that government plays an essential rule in this respect as an enabler rather than a controller. Financial support from all levels of government offices should be rightly focused, catalytic in nature with no distortion of the market, and costs and benefits truly reflected in economic values.

Traditional energy resources in many rural areas have been considered free. This concept does not only undermine the importance of energy efficiency,

but also hinders the promotion of clean fuels, such as biogas and producer gas from gasification. Commercial utilization of biomass resources will be an important step toward better resource management, which is essential for sustainable rural development.

With the increasing gap in income and living standards among different rural regions, energy demand varies greater and greater from area to area. Income for farmers without any other source of income, accounting for 68% of total rural households, has been decreasing in 1999 and 2000, while the national average of rural income has increased 2%. Lower income households, most likely, will still rely on traditional biomass fuel, whereas, higher income farmers will demand modern and clean energy services. Rural energy policy and technology will need to be diverse and locally customized.

SEE ALSO THE FOLLOWING ARTICLES

Biomass Gasification • Biomass: Impact on Carbon Cycle and Greenhouse Gas Emissions • Coal, Fuel and Non-Fuel Uses • Development and Energy, Overview • Indoor Air Quality in Developing Nations • National Energy Policy: China • Rural Energy in India • Sustainable Development: Basic Concepts and Application to Energy • Women and Energy: Issues in Developing Nations • Wood in Household Energy Use

Further Reading

- Byrne, J., Chen, B., and Wallace, W. (1998). Economics of sustainable energy for rural development: A study of renewable energy in rural China. *Energy Policy* 26(1), 45–54.
- China Environment and Development Review (2001). China New and Renewable Energy Network, www.newenergy.org.cn/.
- China Rural Energy and Its Market Potentials, www.scn.cn.net/new_page_70.htm. (Accessed in 2002)
- China Rural Energy Statistics, Ministry of Agriculture, 1997 (1998–1999).
- Commercialization of Solar PV Systems in China, Center for Renewable Energy Development, Energy Research Institute and the U.S. Department of Energy (1996).
- DOE/EIA and OECD data tables, www.geocities.com/~combusem.EIA China Country Analysis Brief, www.eia.doe.gov/emeu/cabs/china.htm.
- McCreary, E. I., Gu, A. Y., Loose, V. W., and Joseph M. Roop, J. M. (1996). China's Energy—A forecast to 2015, Los Alamos National Laboratory.
- Rural Energy Conservation Technologies, Bailiang Zhang (1999). Statistical Abstract. www.stats.gov.cn/ndsj/zgnj/2000/G01c.htm (-G13c), 2000.
- Second Analysis Confirms Greenhouse Gas Reductions in China, www.nrdc.org/globalWarming/achinagg.asp.

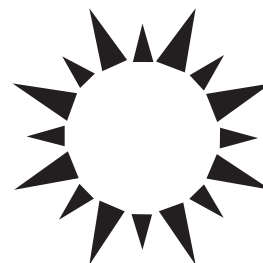
- Wang, J., and Zhao, Z. (2001). "China Energy Development Report."
- Wang, X.D., and Smith, K. R. (1998). Near-term Health Benefits of Greenhouse Gas Reductions: A Proposed Method for Assessment and Application to China, WHO/EHG/98.12.
- Wood Energy Basics, Regional Wood Energy Development Programme (RWEDP) of FAO/UN, www.rwedp.org, 2000.
- Wood Energy, Health and Environment, RWEDP Field Document 58, 2000.
- World Energy Assessment, UNDP (2000).
- Energy for the poor, DIFD (2002).
- Zhang, J., *et al.* (1999). "Greenhouse Gases and Other Airborne Pollutants from Household Stoves in China: A Database for Emission Factors, Atmospheric Environment."
- Zhou, F., and Wang, Q. (2000). "China Energy 50 Years."

Rural Energy in India

IBRAHIM HAFEEZUR REHMAN

The Energy and Resources Institute

New Delhi, India



-
1. Energy Consumption
 2. Rural Electrification
 3. Subsidy Regime for Facilitating Access and Equity
 4. Role of Renewables in Facilitating Transition and Decentralization for Remote Areas
 5. Conclusion

Glossary

administered price mechanism A government-run system of price subsidies aimed at making household energy more affordable.

commercial energy Typically this refers to fossil fuel-derived fuels, such as liquefied petroleum gas, kerosene, natural gas, and coal.

rural electric cooperatives (RECOs) Public-private institutions whose mission is to extend the network and provide services in villages.

traditional energy Typically this refers to biomass-based energy sources, such as fuelwood and animal manure.

India ranks sixth in the world in terms of energy demand, accounting for 3.5% of global commercial energy demand in 2001. India's energy sector is administered at the apex level by five ministries (including the Department of Atomic Energy), whereas the Planning Commission carries out the energy planning exercise for different sectors in the form of 5-Year Plans. Such plans have advocated a strategy focusing on rationalizing tariff, giving a greater role to market forces, enhancing private sector participation, and promoting renewable sources, among other strategies. The Ninth 5-Year Plan (1997–2002) stressed the need for research and development in the fields of energy conservation and demand-side management. It also focused on enhancing the pace of exploration and development in the hydrocarbons sector to reduce dependence on imports. A major thrust was placed on infrastructure

development, particularly energy and power, with plans for enhancing related public expenditure. By the end of the Ninth Plan, the share of commercial energy was expected to increase from 66 to 75%, and the share of noncommercial energy was expected to decline from 34 to 25%. The main elements of power sector reforms in the Tenth Plan (2002–2007) include attempts at rationalization of power tariffs, unbundling of state electricity boards (SEBs), gradual privatization of distribution, and 100% metering of all consumers. In addition, it focuses on hydroelectricity and the possibility of expanding nuclear energy for electricity generation. The implementation of these plans is the responsibility of respective ministries that coordinate with the state governments; the National Development Council provides a forum for coordination at the highest levels of government. In terms of rural access, the more than 700 million rural poor in India have limited access to energy sources, such as electricity, and nontraditional cooking fuels, such as liquefied petroleum gas (LPG). Consequently, the rural energy scenario in India is dependent more on nonmonetized, biomass resources than on monetized or commercial forms of energy. This is amply borne out by the fact that over 72.3% of Indian households depend on traditional sources of energy, such as fuelwood, animal dung, and crop residues, for meeting their cooking and heating requirements. In this regard, over the years, the government, through its various committees and groups, such as the Fuel Policy Committee, the Working Group on Energy Policy, the Advisory Board on Energy, and the Energy Demand Screening Group, has formulated programs aimed at rural energy provision. The efforts have been primarily in the form of national programs for rural electrification, biogas development, improved cookstove dissemination, and promotion of renewable energy technologies (RETs), such as solar cookers and lighting devices. However, despite the existence of these programs for nearly two decades, their impact on the rural energy scenario has been limited.

1. ENERGY CONSUMPTION

In the context of energy consumption, considerations of equity have several dimensions. The most important of these is that adequate energy of good quality is available and accessible to all, such that the poor are able to meet their needs, particularly their needs for cooking fuel and basic lighting. In developing countries such as India, the aspect of quality is important because the use of biomass fuels with low calorific value has serious health implications on account of resultant high levels of indoor air pollution. Clean and good-quality fuels, therefore, need to replace the large-scale use of inferior fuels that are prevalent in India.

The per capita energy consumption in India stands at 479 kg of oil equivalent, which is only 20% of the global average in 1999. One of the main reasons for this is the fact that existing rural energy supply programs often bypass the rural poor because they are unable to interpret their needs and priorities. Even programs specifically meant for the rural poor—supply of kerosene through the public distribution system (PDS) or improved cookstove programs—do not reach them properly or are rejected by them because the initial design does not take into account their needs and priorities.

1.1 Electricity Consumption

With regard to electricity, the consumption pattern was 32% for agriculture, 31% for industry, and 20% for domestic use, and the remaining uses, including commercial and street lighting, make up the final 17%. At the national level, electricity and kerosene are the primary sources for lighting, in both rural and urban areas, and accounted for 99% of households. There has been an increase of 6.5% (from 37 to 43.5%) in the proportion of rural households using electricity as the major source of lighting since 1993–1994. On the other hand, there was a decrease in the percentage of households using kerosene as the primary source of energy for lighting, from 62 to 55.6% in rural India (Fig. 1).

1.2 Consumption of Biomass Fuels

Biomass fuels provide 85–90% of domestic energy and 75% of all rural energy (Table I). Of the total rural energy consumption, approximately 52.5% is met from fuelwood, and fuelwood consumption in 2001–2002 was estimated at 223 million tons (MT).

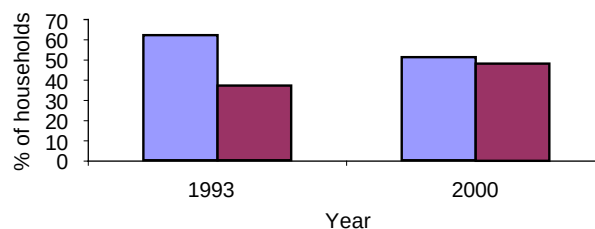


FIGURE 1 Percentage of rural households in India using electricity and kerosene for lighting.

TABLE I

Consumption of Traditional Fuels

- Biomass fuels provide ~90% of the energy in the rural domestic sector and 65% of all rural energy
- Fuelwood is the largest contributor to energy use in rural areas, followed by animal dung and crop residues
- Women and children collect most of the fuelwood, in the form of twigs and branches
- Inefficient use of biomass in traditional devices has serious global and local environmental effects
- Traditional fuels are responsible for adverse impacts on health, especially the health of women and children

Source. Reprinted from National Sample Survey Organization (2001).

Of this, 180 MT was for rural households and the balance was for cottage industries, large hotels, etc. However, against this demand the annual sustainable yield of fuelwood from different land sources is only approximately 86 MT. The consumption of animal dung and agro-waste is estimated at 130 MT; this does not include wet dung used for biogas plants.

The 55th Round of the National Sample Survey Organization survey indicated that firewood and chips were the most important source of energy used for cooking in rural India, with 76% of households dependent on it in 1999–2000. The survey also revealed that only 3% of households have switched from this source to other sources of energy since 1993–1994. The Rural Energy Database compiled by The Energy and Resources Institute has also indicated the highest per capita per day consumption for fuelwood (Fig. 2).

Among rural households, the use of fuelwood and chips declines and the use of kerosene and LPG increases with rising incomes. However, the transition across income classes is not significant, with the result that even in the highest income category, over 70% of rural households depend on fuelwood chips or dung cakes for cooking.

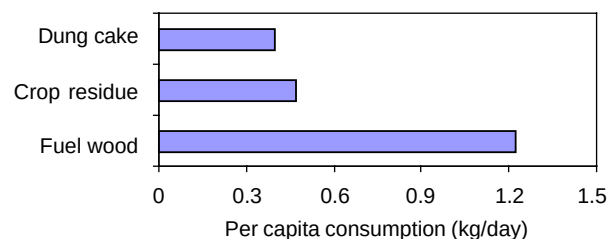


FIGURE 2 Daily per capita consumption of biofuels in rural households in India. Reprinted from TERI (2000), with permission.

There are regional variations in biomass consumption, borne out by the fact that the consumption of fuelwood ranges from 2.04 kg per household per day in Haryana to 14.14 kg per household per day in Orissa. The fuel mix also varies from region to region depending on the resource endowments. Fuelwood consumption is high in states that have considerable forest cover. For instance, in Arunachal Pradesh, it is 10.4 kg per household per day, and in Manipur, it is 10.34 kg per household per day. On the other hand, consumption of dung cakes is high in states such as Uttar Pradesh (4.95 kg per household per day) and Rajasthan (3.41 kg per household per day), which have little biomass cover. Crop residues, being inferior in quality, are used in most areas as backup fuels when there is a shortage of other fuels.

1.3 Commercial Fuel Consumption

Over the years, the dependence of rural households on dung cakes, coke, and coal has fallen and dependence on kerosene and gas has risen. However, in rural areas, the overall shift to commercial fuels has been marginal. Among commercial fuels, kerosene is used predominantly for lighting. Estimates put total kerosene consumption for rural lighting at close to 7 MT. Subsidies on petroleum products, such as kerosene, have contributed to an increase in the demand for kerosene. However, despite a significant increase in the supply of commercial energy, the consumption of commercial fuels, such as LPG, is still negligible in rural areas, with less than 5% of the households using the fuel for cooking.

1.3.1 Kerosene

Kerosene is the main source of lighting in rural India. The total kerosene consumption in India in 2000–2001 was estimated at 10.5 MT, over 50% of which was in the rural areas. In India, there are two systems for the supply of kerosene to households. One is

through the PDS at subsidized rates, forming the main source of supply for rural consumers. The other is the parallel marketing system at market-determined prices. Traditional kerosene lighting devices are used by 74.62 million households throughout India. As is well known, the fuel consumption of these devices is much too high and the quality of light is very poor. Furthermore, their use has adverse impacts on the health of people. Hence, nonelectrified rural households and a majority of electrified households continue to use kerosene for lighting.

The per capita allocation (PCA) of kerosene for states ranges between 10 and 24 kg/year. In 1999–2000, the maximum PCA of kerosene had been frozen at 24 kg/year. Moreover, a private party can import kerosene after obtaining a rating certificate from an approved rating agency. The kerosene is to be sold at market-determined prices by the private parties through their own distribution networks. The quantity of kerosene marketed under the parallel marketing scheme in 1999–2000 accounted for approximately 10% of the total kerosene consumption in the country. The contribution of kerosene through parallel marketing in rural areas, therefore, remains very small.

1.3.2 Liquefied Petroleum Gas

Four public sector undertakings—Bharat Petroleum Corporation Limited, Hindustan Petroleum Corporation Limited, IBP Company Limited, and Indian Oil Corporation Limited—are engaged in the marketing of LPG in the country. With the increased availability of LPG, the number of customers enrolled by these companies has also been increasing. Consumption has grown from 100 MT in 1985 to over 6000 MT in 1999. The number of LPG customers served by the four oil companies as of January 1, 2000 was 43.6 million. However, the rural penetration during this period was just over 1% of the total households.

It is also estimated that in the next decade there will be a rapid increase in the consumption of LPG. Given the near-saturated scenario of the urban market, a large portion of this increase in estimated consumption will occur in rural areas. This is due to the fact that with the increasing nonavailability of traditional fuels, and LPG being a convenient and safe cooking fuel, the government has decided to commence marketing of LPG in rural areas. Furthermore, with a view to penetrate the rural areas, the Ministry of Petroleum and Natural Gas approved a special scheme in July 1999 for the release of new LPG connections in the rural sector against the

surrender of corresponding kerosene quota through the following strategies:

- Opening of extension counters in rural areas within a 15 km radius of the normal trading area of existing distributors operating in the urban areas.
- Award of interim LPG distributorships to the state civil supplies corporation or the concerned state government body until a regular distributor is approved.

In November 1997, the government decided to dismantle the administered price mechanism in a phased manner. As a result, the subsidy on domestic LPG was decreased to 15% in 2000–2001. Moreover, in order to increase LPG availability, the private sector has been allowed to participate in the schemes of parallel marketing of LPG since April 1993. Under the scheme, a private party can undertake the import of LPG and the procedure is the same as that for kerosene. As of January 1, 2000, 695.1 MT of LPG had been imported by the private sector under the scheme.

Despite the fact that LPG has been available for a number of years, it has not penetrated the rural areas of the country. The continued subsidy on LPG has basically hampered the growth of the parallel market and, thereby, adversely affected the penetration in rural areas. On the other hand, the low-income levels in rural areas are responsible for low buying capacities of the rural communities. Moreover, people in the rural areas are not shifting to LPG, as fuelwood is cheap and easily available. For LPG, delivery is not to their homes and means of transportation are not available. Also, there is a fear that LPG is a dangerous fuel. In most companies, the policy is that cylinders are refilled after 21 days. However, owing to logistic and economic reasons, rural customers usually take 3 months to get the cylinders refilled.

1.3.3 Consumption for Agriculture

The agriculture sector is the second largest energy consumer in rural India. Here, the animate energy (human and draught power) accounts for a substantial portion of the total energy consumed. However, the share of human labor and animal power in total energy consumption for agriculture has decreased over the years, indicating progress in farm mechanization. Agriculture accounts for approximately 32% of the total electricity and 95% of the total water consumption in India. Diesel oil and electricity are the major sources of energy used for irrigation. Very small numbers of solar pumps and wind-energy-based pumps are also used.

With the introduction of Green Revolution technologies, the demand for irrigation increased. To meet this increased water requirement, rapid farm mechanization programs along with policy measures such as underpricing of power and irrigation were initiated. As a result, the number of pump-sets has increased and was reported to be close to 20 million, of which just over 12 million had been energized. In 1999–2000, the average farm power availability was 1.15 kW/ha, whereas for desired levels of productivity and appropriate postharvest technologies, a minimum essential estimate is 2 kW/ha.

As mentioned earlier, irrigation is the most important end use, but owing to subsidies, the marginal cost of pumping water is zero. This leads to energy wastage, overpumping, and inefficient crop selection. Flat-rate pumping also masks the true cost of power to farmers. One study indicated that significant energy losses are associated with the distribution of electricity and in the poor selection, installation, maintenance, and operation of electrical motor pump systems.

Suboptimal allocation of resources toward the distribution system places a further burden on the already vulnerable system, causing deterioration in the quality and reliability of supply, leading to further burnout of motors and problems of overloading. The situation is further compounded by the fact that there are a number of illegal connections and cases of cheating in terms of practice of “motor nameplate switching,” whereby higher capacity motors have their labels altered to indicate lower ratings. In a flat-rate tariff system, this further robs the utility of its legitimate revenue.

2. RURAL ELECTRIFICATION

Rural electrification received attention only after the formation of the SEBs in 1948. The Rural Electricity Corporation (REC) was formed in 1969 to administer the Central Plan outlay and provide loans to the SEBs and rural electric cooperatives (RECOs) for rural electrification. The RECOs were entrusted with the task of extending the network and providing services in villages. Thus far, 41 RECOs spread throughout 12 states have been sanctioned by the REC, of which 33 are still in operation and 8 have been taken over by SEBs. The important players in the power infrastructure scenario in India are the SEBs, Power Grid Corporation, Power Finance Corporation, public and private power-producing utilities such as the National Thermal Power

Corporation, Central Electricity Authority, and electricity regulatory commissions. All of the above entities play a very important role in the electrification of rural areas.

Of the 587,258 villages in the country (1991 census), 507,451 (86.4%) were declared electrified by December 31, 2000. Of the over 80,000 nonelectrified villages, most (99.14%) are in Uttar Pradesh, Bihar, Orissa, West Bengal, Meghalaya, Arunachal Pradesh, Tripura, and Rajasthan. The Tenth Plan envisages covering approximately 62,000 of the said villages and the remaining 18,000 remote villages are proposed to be covered by renewable and decentralized sources of energy.

On the other hand, although the per capita consumption of electricity in India was 360 kWh in 1998, the demand for electricity has been increasing at a faster rate, resulting in energy shortages. In order to further the efforts of extending the grid to rural areas, the Government of India launched special programs for rural electrification. One such program was the Minimum Needs Programme, introduced in 1974 with the objective of establishing a network of basic services and facilities in all areas within a given time frame. One of the basic services identified was rural electrification. Earlier, a 100% loan was released to the SEBs under this scheme, but since 2000–2001 it has been released to the state governments. Also, approximately 3.8 million households have been provided single-point light connections under the Kutir Jyoti scheme (1988–1989) for poor rural households.

A 7-year plan is being implemented for rural electrification and has been made part of the minimum needs program; funds are available under the Pradhanmantri Gramodaya Yojana. Under this scheme, 50% of the money is available in the form of grants. In the 2002–2003 budget, interest subsidy has also been provided to states borrowing under this scheme. Money can also be drawn through the Rural Infrastructure Development Fund. Moreover, the REC is also working on a scheme for electrification of 0.4 million hamlets. In addition, discussions are under way with the Ministry of Nonconventional Energy Sources regarding the electrification of 18,000 villages, which cannot be reached through the grid. These villages would be provided electricity through renewables.

However, a disturbing trend is that the rate of electrification of villages has been steadily declining after the Seventh Plan. State agencies are not keen on investing in areas of village electrification and pump-set energization. They do not consider them remunerative. Also, with increasing subsidies, the returns

on the investment affected the loan-repaying capacity of the SEBs.

As mentioned earlier, 86% of the villages have been electrified. The remaining villages are in six states. However, there has been an 85% reduction in the outlay for rural electrification in the Ninth Plan as compared to the Seventh Plan. The situation is further compounded by the definition of village electrification as adopted by the Government of India: “A village will be deemed to be electrified if electricity is used in the inhabited locality within the revenue boundary of the village for any purpose whatsoever.” Therefore, the electrification of a village does not imply that all the houses in a village have electricity. Moreover, long and cumbersome procedures for obtaining connections and paying bills and the distance to repair facilities were some of the factors that affected acceptability. It has also been observed that, starting from the electrification of a village to the procurement of a connection for a household, the time factor can vary from a couple of months to a few years.

Finally, rural India is characterized by small human settlements and given the fact that the energy requirements of these settlements are much lower than in urban and industrial centers, there is high cost of transmission along with severe transmission and distribution losses. Also, although close clusters characterize rural settlements in the plains, settlements in the hilly and forest regions are highly dispersed. Given such a sparse nature of the demand, it becomes economically infeasible to provide power transmission lines to a few houses with a load of a fraction of a kilowatt. Consequently, rural electrification programs have proved expensive and, with the existing constraints on incomes, customers have just not been able to pay for electricity supplies received by them.

3. SUBSIDY REGIME FOR FACILITATING ACCESS AND EQUITY

One of the fundamental causes of rural poverty is inaccessibility to the minimum amount of energy to meet people's basic needs. Hence, the decision makers and planners must assign a high priority to energy provision, irrespective of communities' ability to bear the costs of supply and delivery. Perhaps one of the most critical issues in rural energy development is the nonmonetized nature of many aspects of the rural economy—in particular, the bulk of energy supplies in the form of wood and other biomass fuels. Owing

to availability of “free” sources of energy, it has not been a priority in terms of expenditure on basic amenities. However, in order to move beyond satisfaction of basic needs toward higher levels of economic productivity does mean the introduction into rural societies of modern energy sources.

Recognizing the need to provide access to energy for the poor, the government has provided subsidized energy through either the PDS or the State Electricity Boards. Typically, these subsidies were provided to the residential and agricultural sectors in the form of LPG and kerosene for the former and electricity for the latter. However, the merits of government subsidies designed to accelerate substitution are debatable. Owing to the price distortions that they produce, it is inevitable that most of the subsidies benefit higher income strata and commercial establishments rather than disadvantaged households.

3.1 Subsidy for Irrigation

The total amount of the subsidy on electricity for agriculture and the rural domestic sector has been increased from 202,099 million rupees to 330,814 million rupees in 1999–2000. Of this, just over 70% is provided to the agricultural sector for meeting its irrigation requirements. In contrast, at the all-India level, only 10–12% of the cost of power supply for irrigation purposes is recovered. Moreover, Gulati *et al.* have estimated that the work of each unit of electricity going to agriculture in terms of its contribution to crop output turns out to be 9 rupees per kilowatt-hour for net output at 1995–1996 prices. The farmer, on the other hand, on an average, pays only 0.2 rupees per kilowatt-hour for power. Consequently, over the years, the impact of the subsidies has been twofold: (1) it has led to financial and technical starvation of SEBs and (2) highly subsidized power has led to the overexploitation and misuse of groundwater.

3.2 Deterioration in the Condition of State Electricity Boards

Owing to poor or negligible revenue, the operating as well as financial condition of most of the SEBs has deteriorated over time. The main cause is their inability to recover costs from different consumer segments. For SEBs to break even, state governments need to make budgetary provisions in excess of 5% of their total revenues. If the SEBs are to earn a commercial rate of return (15–16%) on their assets, the subsidies must be five times more than those needed to break even, so that the required subsidy, in

1999, would equal 25% of the state government revenues, or 12 billion dollars.

3.3 High Rate of Groundwater Withdrawal

The rate of groundwater withdrawal is closely linked to energy prices. In most states, power to agriculture is supplied at a flat rate based on the capacity (horsepower) of the pump rather than on the actual quantity of electricity consumed to pump the water. Thus, a high implicit subsidy is built into the system. As a result, many regions in India are witnessing shortages in water supply, with groundwater levels having fallen as much as 1 to 3 m per annum to levels more than 70 m or more below their levels 30 years ago. Approximately 12% of all aquifers are severely overdrawn. Overexploitation of groundwater by farmers has lowered water tables. Overexploitation is also a significant contributor to India's growing carbon emissions, as considerable additional pumping energy is required to extract deeper water supplies.

3.4 Subsidy on Liquefied Petroleum Gas and Kerosene

Even after the dismantling of the APM, it was decided that the government would continue to subsidize kerosene and cooking gas. Consequently, in the 2002–2003 union government budget, the government has set aside 64.95 billion rupees as a subsidy on kerosene and cooking gas for fiscal year 2002–2003.

4. ROLE OF RENEWABLES IN FACILITATING TRANSITION AND DECENTRALIZATION FOR REMOTE AREAS

India is endowed with abundant natural and renewable resources of energy. The country has been able to achieve significant capacity addition of 1367 MW through wind farms and ranks fifth in the world. Overall, the renewable energy sector contributes over 3400 MW, supplementing and complementing the conventional energy sector in meeting the country's growing energy needs. Moreover, it also helps in meeting the global demand for limiting the emissions of greenhouse gases into the atmosphere. The status of the installation of RETs is indicated in Table II.

TABLE II
Status of Installation of Renewable Energy Technologies

Technology	Potential	Achievement	India's position (in the world)
Biogas plants (units)	12 million	3.3 million	2nd
Improved cookstoves (units)	120 million	34 million	2nd
Wind (MW)	45,000	1702	5th
Small hydro (MW)	15,000	1463	10th
Biomass power/cogeneration (MW)	19,500	468	4th
Biomass gasification (MW)		53	1st
Solar photovoltaic (MW/km ³)	20	107	3rd
Waste to energy (MW)	2,500	25	—
Solar water heating	140 million km ² collector area	0.68 million km ² collector area	—

Source. Reprinted from Ministry of Nonconventional Energy Sources (2003).

By their very nature, rural areas lend themselves amenable to the use of renewable energy. RETs generally provide decentralized and modular ways of meeting energy demands. Rural areas in India have an abundance of solar energy that can be exploited for both thermal and electrical applications. Biomass energy is conventionally referred to as traditional energy but, when used in modern technology applications, also qualifies as renewable energy. The dissemination of RETs in the country has been mainly through government programs that have primarily helped in introducing rural people to the technologies, but have not been successful in wide-scale dissemination and penetration. Given the nonviability of extending the power grid to remote areas, RETs can provide a means by which a decentralized supply of energy for such areas becomes feasible and close to being economically viable. In this regard, the plans of the concerned ministry are indicated in Table III.

Clearly, most of these initiatives are geared to meet rural energy requirements. In essence, India is on the edge of a new revolution in the management of rural energy problems and must, therefore, devise policies that move away from the mindsets of the past and build on the technologies for the future.

5. CONCLUSION

Since a major part of the planning process pertaining to rural energy has been left to consumers themselves, the cost of the fuel, economic status of the

TABLE III
Renewable Energy Plan 2012 of the Ministry of Nonconventional Energy Sources, Government of India

- Coverage of 30 million households through improved cookstoves
- Setting up of an additional 3 million biogas plants
- Deployment of 5 million solar lanterns and 2 million solar home systems
- Deployment of solar water heaters in 1 million homes
- Addition of 10,000 MW of renewables-based power
- Electrification of 18,000 remote villages

Source. Reprinted from The Energy and Resources Institute (2002).

community, and easy availability are the determinants for the type of fuel and technology used for meeting energy needs. One of the main reasons for the absence of a modern energy supply in rural India is the low purchasing power of the people. Given the fact that a vast majority of rural people have limited cash, they find it difficult to afford the initial costs of acquiring more efficient and cleaner fuels and technologies. Consequently, the rural poor are drawn into a cyclic pattern of energy consumption that adversely affects the resource base as well as the individuals.

With the emphasis on electricity reforms, accelerated rural electrification, and increasing the share of renewables in the overall energy consumption, the scenario is changing. The country is on the threshold of a major change in the energy supply-and-demand scenario and the right policies and initiatives will facilitate the realization of the objective of energy for all.

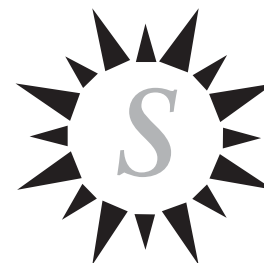
SEE ALSO THE FOLLOWING ARTICLES

Biomass Gasification • Biomass: Impact on Carbon Cycle and Greenhouse Gas Emissions • Development and Energy, Overview • Indoor Air Quality in Developing Nations • National Energy Policy: India • Rural Energy in China • Sustainable Development: Basic Concepts and Application to Energy • Women and Energy: Issues in Developing Nations • Wood in Household Energy Use

Further Reading

- Commissioner, Registrar General of the Census. (1991). "Census of India, 1991, Table: Household Amenities and Assets." Registrar General of the Census, Commissioner, New Delhi, India.
- Ministry of Petroleum and Natural Gas. (2000). "Annual Report 2000–2001." Ministry of Petroleum and Natural Gas, Government of India, New Delhi, India.
- National Sample Survey Organization 55th Round. (2001). "Energy Used by Indian Households 1999–2000," Report 464(55/1.0/6). National Sample Survey Organization, Ministry of Statistics and Programme Implementation, Government of India, New Delhi, India.
- Natrajan, I. (1985). "Domestic Fuel Survey with Special Reference to Kerosene," Vol. 1. National Council for Applied Economic Research, New Delhi, India.
- Padmanaban, S., and Sarkar, A. (2001). "Electricity Demand Side Management (DSM) in India: A Strategic and Policy Perspective." Office of Energy, Environmental, and Enterprise, USAID, India. Presented at the International Conference on Distribution Reforms, October 12, 2001.
- Planning Commission. (2001). "Annual Report on the Working of the State Electricity Boards and Electricity Departments: Power and Energy Division." Planning Commission, Government of India, New Delhi, India.
- Planning Commission. (2002). "Tenth Five-Year Plan: 2002–2007, Vol. II, Thematic Issues and Sectoral Programmes." Planning Commission, Government of India, New Delhi, India.
- Ravindranath, N. H., and Hall, D. O. (1995). "Biomass, Energy and Environment: A Developing Country Perspective from India." Oxford University Press, New York.
- The Energy and Resources Institute (TERI). (1995). "Environmental Considerations and Options in Managing India's Long Term Energy Strategy (Eco-miles)." TERI, New Delhi, India.
- TERI. (2000). "Rural Energy Matters: The Dhanwas Experience," p. 184. TERI, New Delhi, India.
- TERI. (2001). "Directions, Innovations, and Strategies for Harnessing Action (DISHA)" (R. K. Pachauri and R. K. Batra, Eds.), p. 310. TERI, New Delhi, India.
- TERI. (2002). "Defining an Integrated Energy Strategy for India: A Document for Discussion, Debate, and Action." TERI, New Delhi, India.

Service and Commerce Sector, Energy Use in



BERND GEIGER and PETER TZSCHEUTSCHLER

Munich University of Technology
Munich, Germany

-
1. Commerce/Services: Element of a Final Energy Balance
 2. Demarcation of Residential and Commerce/Services
 3. Commerce/Services/Handcraft/Agriculture
 4. Structure Analysis of the Sector Commerce/Services in Germany
 5. Summary

Glossary

commerce/services sector Contains commerce, handcraft, construction, public and private services, trade, agriculture, and others.

final energy Supplied to the consumer to be converted into useful energy (e.g., electricity at the wall outlet).

residential sector Comprises private households.

The world has to deal with an increase of the global energy consumption. This affects each of the classical consumer sectors: industry, transportation, residential, and commerce/services. To a large extent, the rising energy demand will be met by extending the use of conventional energy carriers such as oil, natural gas, and coal. This is a global trend, even though the characteristics of this development may vary from continent to continent and from country to country. To detect the reasons for energy use, inventories of energy consumption and analyses of energy application were always practical. Across-the-board comparisons of continents or countries are of little help. Taking into account parametrically effective factors on the energy consumption brings some transparency and understanding of what would otherwise be rated and commented falsely. This applies in particular to the commerce/services sector, which had occupied a shadow existence in the national energy balances for a long period. Fortunately, the provision of data to this sector has

improved substantially, and many gaps can be closed. Creating a first country-wise data framework for energy consumption, which can be consulted for comprehensive country-related analyses, is a large merit of international statistics. However, this may not lead to a blind belief in the reliability of energy consumption data. Critical notes are therefore always helpful and appropriate. This article deals with the energy consumption structure of the commerce/services sector, indicates parameters of influence and supplies results and perceptions that apply worldwide, for chosen countries, and in Germany.

1. COMMERCE/SERVICES: ELEMENT OF A FINAL ENERGY BALANCE

This article examines the significance and importance of the commerce/services sector by consulting the energy consumption data for 1995, as published in 1998 by the World Energy Council. Based on the six continents—North America, Central/South America, Europe, Africa, Asia, and Oceania—continental energy balances were set up as shown in Fig. 1. It appears that the proportion of the industry is situated between 18% and 40%, the transportation sector accounts for 22% to 45%, private households account for between 13% and 58%, and the proportion of the commerce/services sector varies from 3% to 12%.

By differentiating between continents, it is possible to mark out climatic influences as well as population density and economic prosperity, as shown in Figs. 2 and 3. As a consequence, a large bandwidth of the specific final energy consumption can be noted. In terms of final energy demand, North America consumes 155 GJ per capita and year, while Africa consumes 16 GJ per capita and year only.

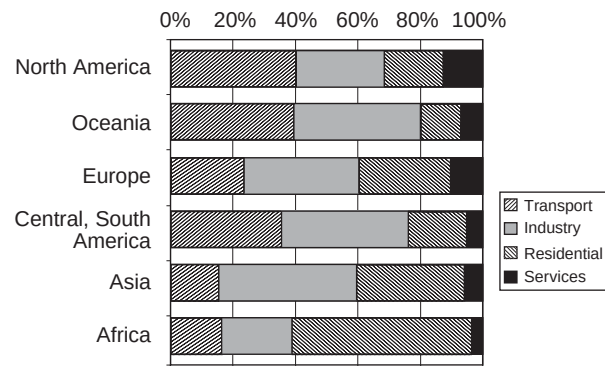


FIGURE 1 Contributions of the different sectors to the final energy demand in 1995; 100% represents the total demand of final energy.

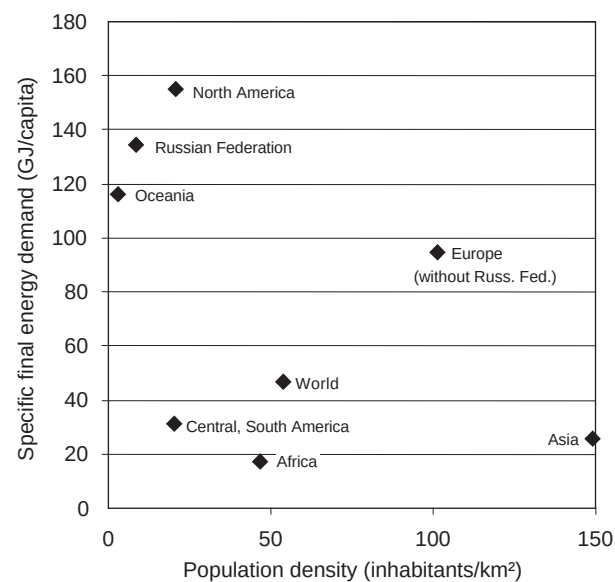


FIGURE 2 Specific final energy demand in relation to the continental population density in 1995.

It can be postulated that an increasing population density reduces the expenditure for transportation and thus its proportion of the final energy consumption. Shorter distances of transport and communication as well as shorter commuter distances result in lower consumption values in the traffic sector (Asia, Europe, Africa). On the other hand, a low population density causes substantial consumption impulses for the traffic sector. This does not remain without effect on the energy consumption in the sector commerce/services.

Each society requires products and services, whereby substantial levelings are to be assumed—not only depending on the prosperity and the gross domestic product, but also on the capability for exports. A

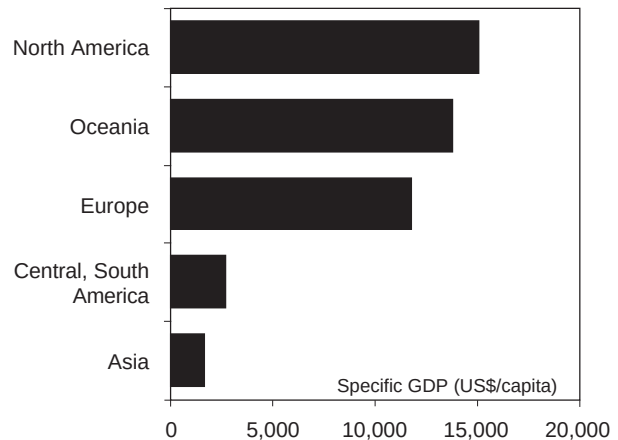


FIGURE 3 Specific GDP of the different continents in 1995.

sharp demarcation of industry and handcraft, respectively, of industrial structures and small trade activities regarding the associated energy consumption causes some problems. Thus, the proportions for industry and commerce/services in Fig 1 should be considered as a reference value instead of as a definite mirror image of accurate allocation.

The common life of each society, whether it is shaped industrially, agriculturally, or mixed, requires services and the supply of everyday products such as nourishments. In the course of social restructuring with increasing prosperity, the tendency toward division of labor is enforced. This means that while the national income is still low, handcraft, small trade, and agriculture are, to a large extent, connected to the residential sector. With increasing prosperity, a separate services sector emerges and organizes. These circumstances are also reflected in Fig 1: while in industrialized continents with above-average per capita income (North America, Oceania, Europe) the activities of commerce and public services are “appropriate,” verifiable, and delimitable to other consumption sectors, this is not always possible in Asia and Africa, where these activities are partly contained in the residential sector.

We are still a long way from an accurate image and sharp delimitations, even though large progress in the collection of international consumption structure data has been made.

2. DEMARCATION OF RESIDENTIAL AND COMMERCE/SERVICES

The boundaries between the residential and commerce/services or handcraft and agriculture are often

diffuse as far as the associated energy consumption is concerned. This becomes relevant whenever the same building is used for residential and occupational purposes, when a household cannot be separated from an agricultural supplementary income, or if members of a family are temporarily bound in a business.

Statistically proven consumption delimitations have to be considered under these aspects too: what is separately accounted for as commerce/services in one country can be accounted for under the category of residential in another country. Therefore, careful handling of sector-specific consumption data is inevitable if these are to be analyzed or interpreted or if a landspreading transverse comparison is to be carried out, as is demonstrated next.

Figures 4 and 5 show the per capita consumption of electricity and fuels for households and commerce/services, again aggregated by continents. It is remark-

able that the electricity consumption of commerce/services is generally higher than the electricity consumption of the residential sector; for residential it ranges from 0.3 GJ/cap in Africa to 10 GJ/cap in North America. The fuel consumption for residential is significantly higher than it is for commerce/services; the values of the latter spread between 0.6 GJ/cap in Africa and 12 GJ/cap in North America.

For a better understanding of these specific consumption values, one should know that from 20 to 50% of the population is involved in the sector "commerce/services/handcraft/agriculture" because of employment, occupation, or simply for covering vital needs.

This note seems necessary, if energy consumption is expressed as giga joule per capita and is not allocated to its origin (e.g., earner or employee of a consumption sector). If earner or employee is chosen as base factor, the specific consumption is between two and five times higher according to the relation of inhabitants per earner/employee.

3. COMMERCE/SERVICES/ HANDCRAFT/AGRICULTURE

Figure 6 shows the final energy demand per capita for the sector commerce/services (handcraft and agriculture might also be avcontained). It is differentiated by continents and therefore also by climates and by consumption of fuel and electricity.

For the fuel consumption, mean values between 10 GJ/cap and 0.15 GJ/cap occur in the following order: North America, 10 GJ/cap; Europe, 5.6 GJ/cap; Oceania, 2.6 GJ/cap; Asia, 0.9 GJ/cap; Africa,

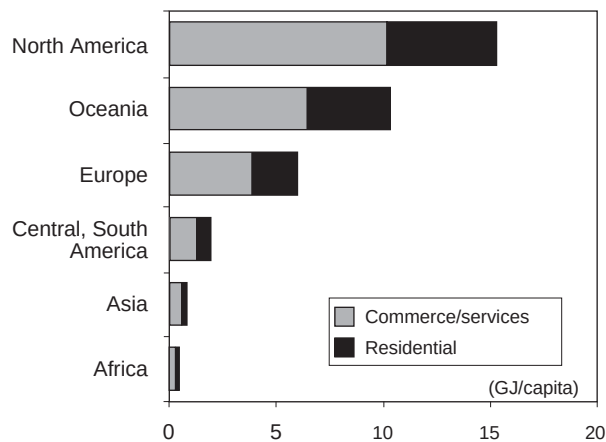


FIGURE 4 Specific electricity consumption of the commerce/services and residential sectors in 1995.

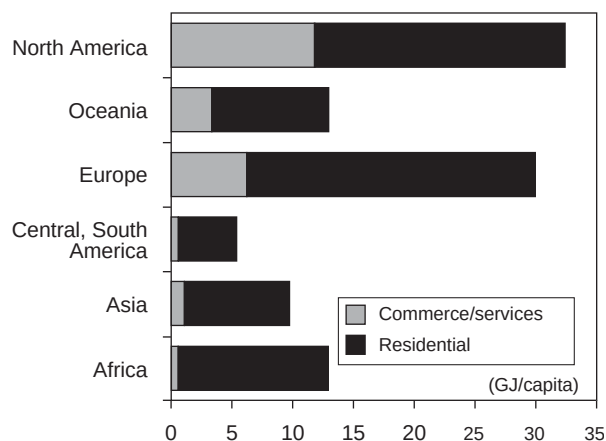


FIGURE 5 Specific fuel consumption of the commerce/services and residential sectors in 1995.

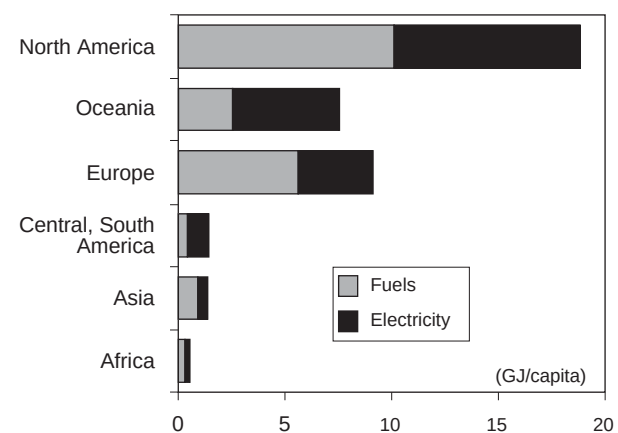


FIGURE 6 Specific final energy demand of the commerce/services sector in 1995.

0.25 GJ/cap; and Central, South America, 0.15 GJ/cap. The energy consumption values are mainly due to heating, conditioning of the work space, and the supply of warm water. The required area per employee at his or her workplace must be considered.

Based on temperature conditions as they occur in North America and Europe, an estimated area requirement of 40 m² per wage earner and a heating requirement of 200 kWh per m² and year and a specific consumption of 8000 kWh or 28.8 GJ per wage earner and year can be calculated. Rated by 0.3 wage earners per capita, an energy consumption for space heating in the sector commerce/services of 8.6 GJ per capita and year can be assumed.

This is a standard consumption level for heating purposes in many prospering industrial societies.

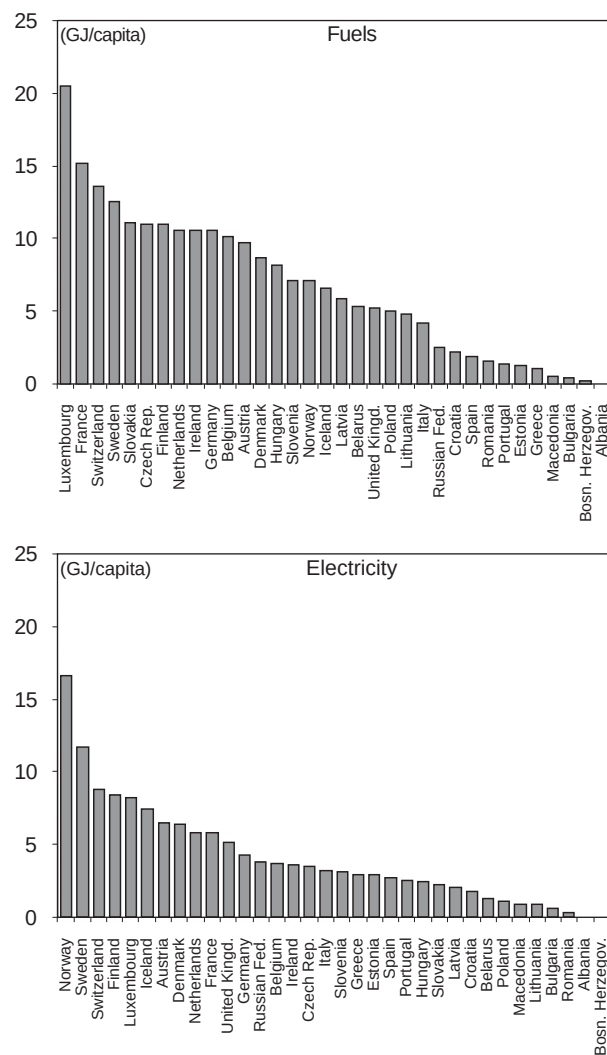


FIGURE 7 Final energy demand of the commerce/services sector in Europe in 1995.

Another standard consumption level is represented by the need for heated water (for washing, bathing, showering), which accounts for 300 kWh per wage earner and year, respectively, 1.08 GJ per wage earner and year. This corresponds to 0.3 GJ per capita and year. Thus, with only a few characteristic values, the specific final energy demand of 10 GJ per capita and year can be explained and analyzed to a large extent.

Finally, area requirements of 40 m² per wage earner correspond to empirical average values, which may vary significantly in individual cases. Employment structures with areas per employee of 100 m² are also not unusual, if attractiveness within the service area and customer demands require this.

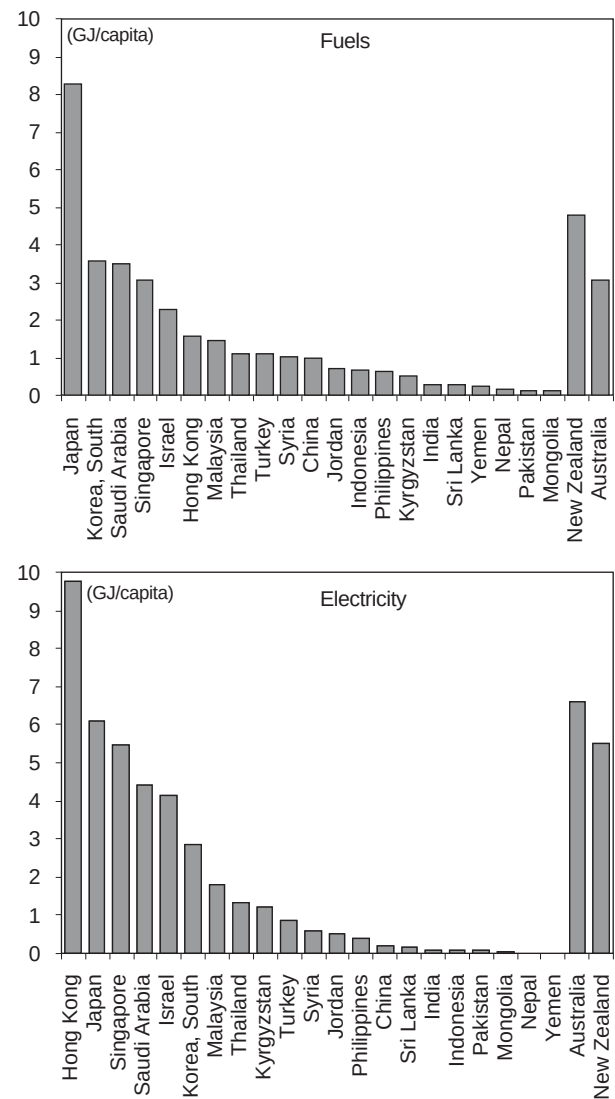


FIGURE 8 Final energy demand of the commerce/services sector in the United States in 1995.

Energy consumption values for room conditioning (heating, ventilation, air conditioning) of $200 \text{ kWh/m}^2 \cdot \text{a}$ are also average values and result in a large bandwidth. On the one hand, thermal protection and heat recovery systems supply appropriate solutions for the reduction of consumption without a loss of comfort. On the other hand, objects with a specific consumption of more than $1000 \text{ kWh/m}^2 \cdot \text{a}$ —no rarity, by the way—are an unnecessary waste of energy.

More details can be derived from Figs. 7 to 10. The European countries indicate a large bandwidth in the specific electricity and fuel consumption, which is due to climatic influences and different levels of prosperity (Fig. 7). This also applies to the

countries in North and South America. Cool Canada and warm South America, rich North America and poor South America, compose basic conditions that influence the specific requirements of electricity and fuel (Fig. 8). The countries of Asia and Oceania are generally situated in warm and hot climates, which is reflected in low specific fuel consumption. While this temperature dependency applies to most of these countries, Japan and New Zealand show an unexpected behavior. In all remaining countries, this factor is temperature dependent as far as possible. The specific electricity consumption represents the prosperity of Hong Kong, Japan, Australia, and New Zealand (Fig. 9). Warm to hot climate zones in Africa lead inevitably to small specific fuel consumption. The economic position of Africa (with a small

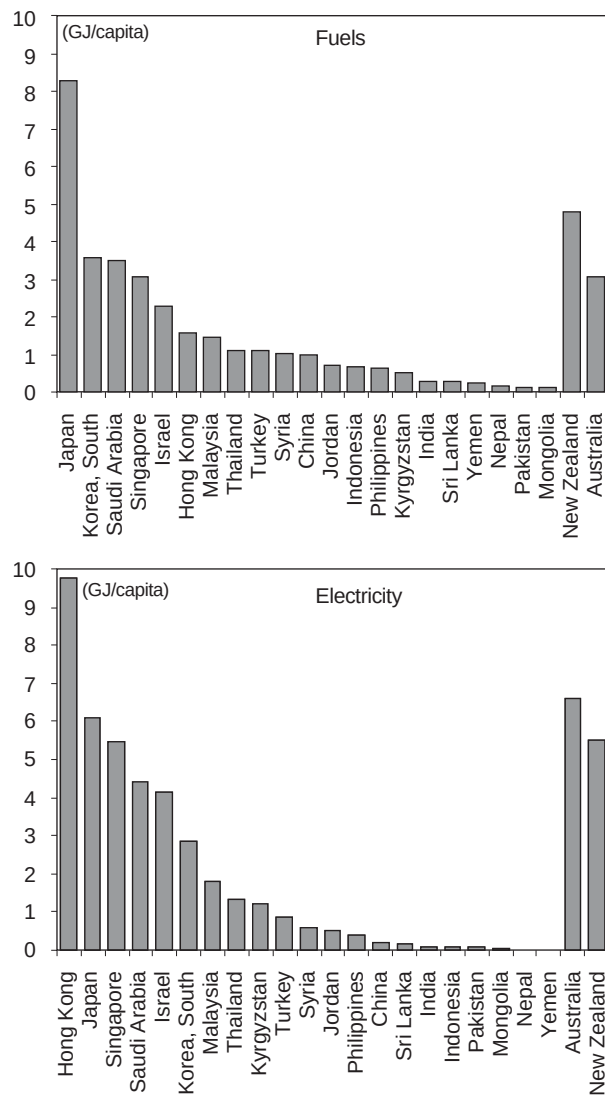


FIGURE 9 Final energy demand of the commerce/services sector in Asia and Oceania in 1995.

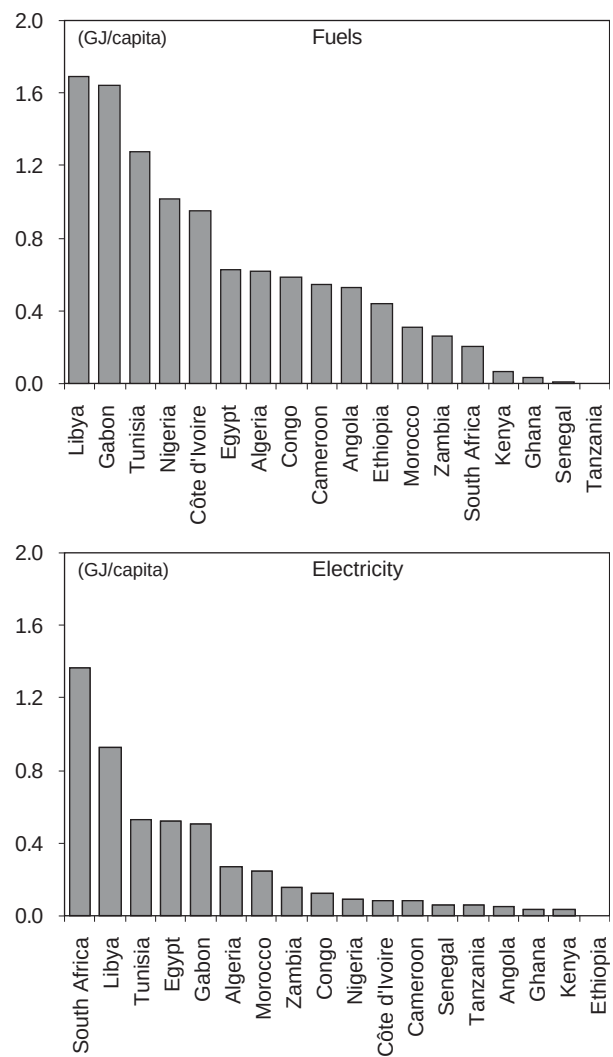


FIGURE 10 Final energy demand of the commerce/services sector in Africa in 1995.

TABLE I

Sociodemographic and Energetic Data of Chosen Countries in 1995

Country	Inhabitants [Mio.]	Average temperature [°C]	Specific GDP [U.S. \$/capita]	Commerce/ services	
				Fuels [GJ/capita]	Electricity [GJ/capita]
Africa					
Algeria	28.1	18.5	2170	0.62	0.26
Angola	11	24	332	0.53	0.05
Congo (Brazzaville)	2.6	25	1076	0.59	0.12
Cameroon	13.3	26	768	0.54	0.08
Egypt	57.5	21.5	884	0.62	0.52
Ethiopia	54.7	16.5	140	0.44	0.00
Gabon	1.1	25.5	6251	1.65	0.5
Ghana	17.7	27	436	0.03	0.03
Côte dl’voire	14.2	26.5	833	0.95	0.08
Kenya	30.5	20.5	303	0.06	0.03
Libya	5.4	20	5282	1.69	0.92
Morocco	26.4	19	1037	0.31	0.24
Nigeria	99	27	329	1.02	0.09
South Africa	39.5	18.5	2814	0.2	1.36
Senegal	8.6	25.5	711	0.01	0.05
Tunisia	9	18	1694	1.28	0.52
Tanzania	28.3	25	144	-	0.05
Zambia	9.1	21	332	0.26	0.15
North and South America					
Canada	29.4	6	12061	23.18	14.38
Jamaica	2.5	27	1799	0.47	1.58
Mexico	90.5	18.5	3132	0.88	0.59
Trinidad and Tobago	1.3	26	4245	0.9	1.02
United States	263	15	23368	14.42	13.06
Argentina	34.8	16.5	5217	0.83	1.15
Bolivia	7.4	13	733	0.27	0.21
Brazil	155.8	22.5	3215	0.35	1.28
Chile	14.2	14.5	3060	0.77	-
Paraguay	4.8	24	1283	0.6	0.53
Peru	23.5	16.5	1885	0.55	0.39
Uruguay	3.2	16	3102	0.61	1.53
Venezuela	21.8	24.5	2611	1.44	2.53
Asia					
Sri Lanka	18.1	27.5	551	0.26	0.16
China	1220.5	13	519	0.97	0.18
China, Hong Kong	6.2	15	15881	1.56	9.77
Indonesia	194.8	27.5	913	0.68	0.1
India	922	27.5	415	0.29	0.1
Israel	5.5	20	12634	2.26	4.15
Japan	125.5	15	25696	8.27	6.1
Jordan	5.7	17.5	1007	0.72	0.5
Kyrgyzstan	4.6	4	311	0.52	1.18
Korea, South	45.1	11.5	8069	3.58	2.87
Mongolia	2.4	-2.5	-	0.11	0.05

continues

Table I continued

Country	Inhabitants [Mio.]	Average temperature [°C]	Specific GDP [U.S. \$/capita]	Commerce/ services	
				Fuels [GJ/capita]	Electricity [GJ/capita]
Malaysia	20.7	27.5	3149	1.47	1.81
Nepal	20.3	18.5	229	0.16	0.01
Pakistan	130.3	25	348	0.13	0.07
Philippines	70.3	27.5	706	0.61	0.38
Saudi Arabia	18.3	27.5	6412	3.51	4.4
Singapore	3.5	27.5	16328	3.07	5.45
Syria	14.2	18.5	1235	1.04	0.58
Thailand	59.4	27.5	1681	1.1	1.34
Turkey	60.6	15	2912	1.09	0.86
Yemen	15.4	29.5	511	0.22	-
Oceania					
Australia	18.1	18	18705	3.07	6.6
New Zealand	3.7	12	13610	4.77	5.49
Europe					
Albania	3.6	16.5	510	-	0.03
Austria	8.1	9.5	21846	9.71	6.44
Belgium	10.1	10	20310	10.15	3.65
Bosnia and Herzegov	4.2	10	309	0.23	-
Belarus	10.3	5	2139	5.32	1.24
Bulgaria	8.4	10.5	2016	0.42	0.62
Denmark	5.2	8.5	31281	8.73	6.31
Ireland	3.6	10	16815	10.57	3.6
Estonia	1.5	5	3033	1.29	2.88
Czech Republic	10.3	8	6531	11	3.5
Finland	5.1	4.5	24646	10.97	8.38
France	58.1	11.5	21678	15.21	5.75
Germany	81.7	9	25753	10.54	4.21
Greece	10.5	17.5	8586	1.04	2.88
Croatia	4.7	15	3846	2.15	1.72
Hungary	10.2	11	2989	8.21	2.44
Iceland	0.3	5	24148	6.56	7.37
Italy	57.3	15	20187	4.21	3.18
Latvia	2.5	6	2570	5.84	2.06
Lithuania	3.7	5	2414	4.84	0.88
Slovakia	5.4	10	2102	11.07	2.25
Luxembourg	0.4	9.5	32898	20.52	8.2
Macedonia	2	15	752	0.53	0.89
Netherlands	15.5	9.5	20364	10.57	5.8
Norway	4.4	5.5	31648	7.11	16.51
Poland	38.6	7.5	1705	4.99	1.09
Portugal	9.9	16	7294	1.31	2.51
Romania	22.7	10.5	1473	1.54	0.27
Russian Federation	148.1	3.5	2428	2.47	3.76
Slovenia	2	10	8492	7.16	3.11
Spain	39.2	15.5	13409	1.91	2.73
Sweden	8.8	6	26544	12.57	11.66
Switzerland	7	10	32386	13.64	8.74
United Kingdom	58.6	10	17640	5.2	5.05

specific national income) is again reflected in the low specific electricity consumption (Fig. 10).

Is it possible to deduce overall relations from the facts, like the influence of climate zones and prosperity on the electricity and fuel consumption? To explore this question, we gathered 93 countries in Table I and used these data as a basis for further analyses. Table I contains data for electricity and fuel consumption, GDP (each referred to the number of inhabitants), as well as climatical information.

Generally the application of electrical energy is favored if a high proportion of electricity is generated by hydropower. Therefore, the following countries are excluded from further analyses: Norway, Sweden, Finland, Iceland, Canada, Switzerland, and France.

In Fig. 11 depicts the relationships between the annual average temperature and the specific electricity and fuel consumption. The higher the annual average temperature, the smaller the influence of fuel consumption. The lower the annual average

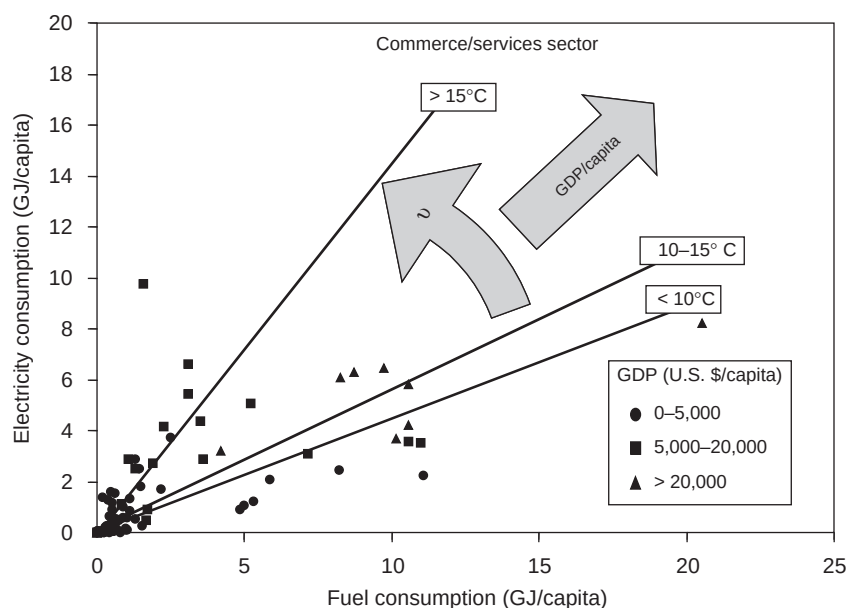


FIGURE 11 Energy consumption in relation to climate and prosperity in 1995.

TABLE II

Consumption Data for Space Heating in Commerce/Services

Country	Floor area in commerce, services per capita m ² /capita	GDP commerce/services US \$/capita	Service sector fuel/heat 1995 GJ/(m ² *a)	Mean temperature °C
Italy	6.5	10,300	0.72	15.0
Japan	10.5	11,000	0.70	15.0
Norway	16.0	11,300	0.58	5.5
Australia	17.5	11,600	0.20	18.0
Denmark	17.5	10,000	0.59	8.5
USA	25.0	16,400	0.60	15.0
Sweden	15.5	10,500	0.81	6.0
Canada	16.0	10,500	0.99	6.0
New Zealand	14.0	9,300	0.35	12.0
Germany	17.8 ^a		0.50	9.0
France			0.20	11.5
Finland			1.05	4.5

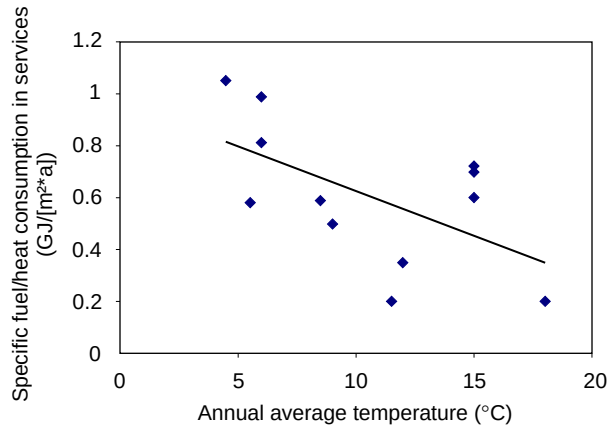


FIGURE 12 Specific fuel and heat consumption of commerce/services in relation to the annual average temperature in 1995.

temperature of a country, the higher the fuel consumption per capita, since space heating gains more and more importance. This behavior sets in from an annual average temperature of 12.5°C. With annual average temperatures between 10 and 0°C, the fuel consumption caused by space heating increases distinctly. Rising prosperity in general leads to a higher specific consumption of electricity and fuel due to more and better electric equipment and heating of larger and well-conditioned rooms.

In Table II presents interesting data for a cross-national comparison (as completed in 2000 by Geiger and Harrison). The average available floor area per capita in the commerce/services sector, GDP, the annual average temperature, and the resulting energy demand for space heating per area and year or per capita and year are included. In Fig. 12 the coherence between cold climate and rising energy consumption becomes obvious too.

4. STRUCTURE ANALYSIS OF THE SECTOR COMMERCE/SERVICES IN GERMANY

Much can be learned from observing the commerce/services sector in Germany. Figure 13 shows the energy consumption broken down into energy carriers and applications such as process heat, space heating, mechanical energy, lighting, and communication for a period of 10 years. By collecting energy requirement statistics, investigating energy consumption, and analyzing changes from year to year, we have come to the conclusion that short-term changes are, in most cases, caused by climatical or cyclical economical reasons fluctuations. Tendentious changes

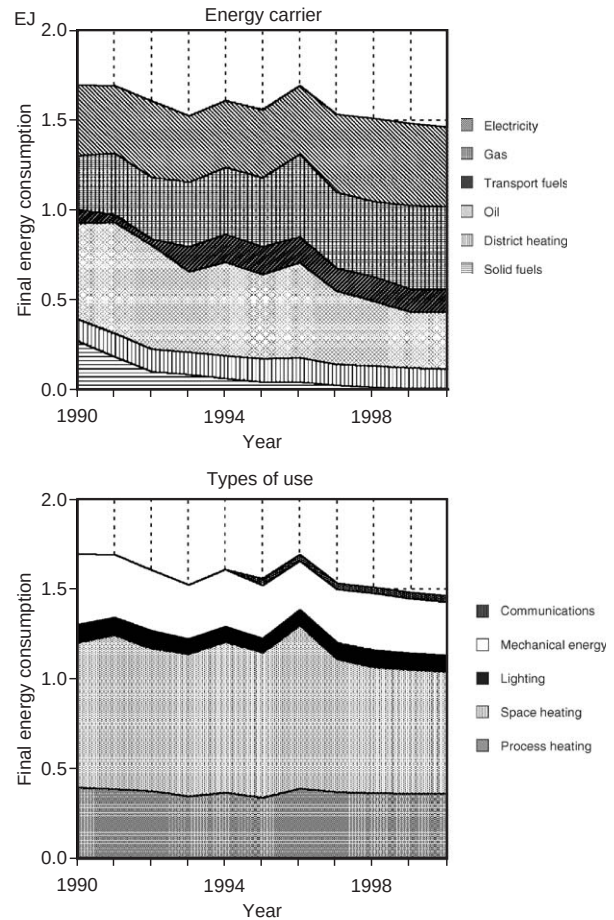


FIGURE 13 Final energy consumption of the commerce/services sector in Germany.

during larger periods result from changes in effective areas, technical equipment, and user requirements.

Beyond this, the question of who needs how much energy and for which purpose is gaining interest. “Who” refers to typical consumer groups from the commerce/services sector. By investigating how much, it is not only of interest to analyze the particular consumption proportions of different consumer groups, but also which consumption per person employed or per m² net area is typical. We select “persons employed” and “net area” as base factors, because these are relatively available and also are usually known. “For which purpose” specifies how the energy under investigation is being applied, such as for lighting, electrical drives, communication, space heating, or process heat.

On the basis of inspections, a metering campaign, and a survey of about 2800 companies and facilities during the 1990s, detailed information on the energy consumption and its key indicators could be sampled,

detected, and measured. Twenty-three consumer groups could be distinguished, and data for an additional nine consumer groups could be retrieved from secondary statistical material, with sufficient information quality. Within the scope of the survey, information on electricity and fuel consumption for

heating and transportation could be collected. On the basis of equipment size, process load, and user profiles, the electricity and fuel consumption could be analyzed and differentiated according to the purpose of use.

Table III shows the calculated overall consumption by small consumers in 1994; additional

TABLE III

Final Energy Consumption: Consumer Groups of Commerce/Services and other Sector, Germany, 1994

Unit	Employees [Mio.]	Electricity [TWh/a]	Fuels and district heating [TWh/a]	Transp. fuels in the sector		
				Commerce/ services [TWh/a]	Transport [TWh/a]	Operating area [Mio. m ²]
Metal Working Industry ^a	0.96	2.33	8.50		10.92	53
Automobile trade and commerce	0.14	0.46	1.96		0.99	18
Timber industry ^a	0.21	0.68	4.25		2.36	22
Printing and paper industry ^a	0.11	0.43	1.35		0.68	6
Bakeries	0.19	1.18	5.99		1.28	7
Butcheries	0.15	1.03	2.13		0.87	6
Other food trade and commerce	0.03	0.09	0.39		0.37	2
Laundries	0.09	0.60	3.37		0.50	3
Horticulture	0.27	0.60	6.90		1.02	35
Retail trade	3.07	12.99	35.50		14.54	205
Wholesale trade	1.57	7.04	16.25		18.57	184
Credit institutes and insurances	1.04	2.80	7.22		4.27	36
Accommodations	0.35	3.25	10.91		1.02	51
Catering	0.86	5.01	11.98		4.32	45
Publishing companies	0.18	0.48	0.70		1.07	6
Other services	4.01	5.44	23.63		31.99	106
Nonprofitable organizations	1.68	4.27	24.48		2.88	152
Public corporations and social security	4.16	5.92	32.07		2.94	153
Hospitals	1.06	4.64	15.36		0.13	47
Schools	1.27	4.94	27.88		0.05	130
Baths	0.07	1.70	5.56		0.00	6
Clothing trade and commerce	0.33	1.97	10.20			
Trading agencies	0.25	1.12	2.59			29
Postal services	0.35	0.57	1.19			6
Telecommunications	0.23	2.98	0.63			
Postbank AG	0.02	0.05	0.14			1
Railway (Deutsche Bahn AG)	0.22	1.78	7.98			
Inland navigation	0.04	0.00	0.00			
Airports	0.06	0.61	0.60		0.02	
Shipping and store house	0.15	0.07	0.15			
Street lighting	0.00	3.33	0.00			
Milit.departments	-	0.81	3.25	11.39		
Main construction trade	3.00	2.47	16.87		32.45	136
Agriculture	0.79	4.90	20.30	25.20		624
Total	26.89	86.55	310.27	36.59	133.22	2.070

^a Less than 20 employees.

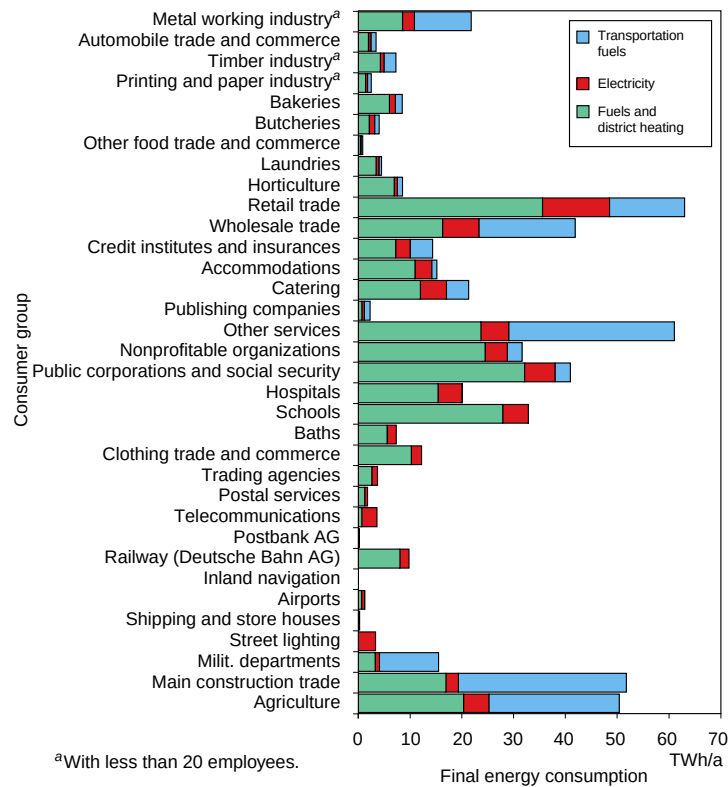


FIGURE 14 Final energy consumption: consumer groups of commerce/services and other sectors; Germany, 1994.

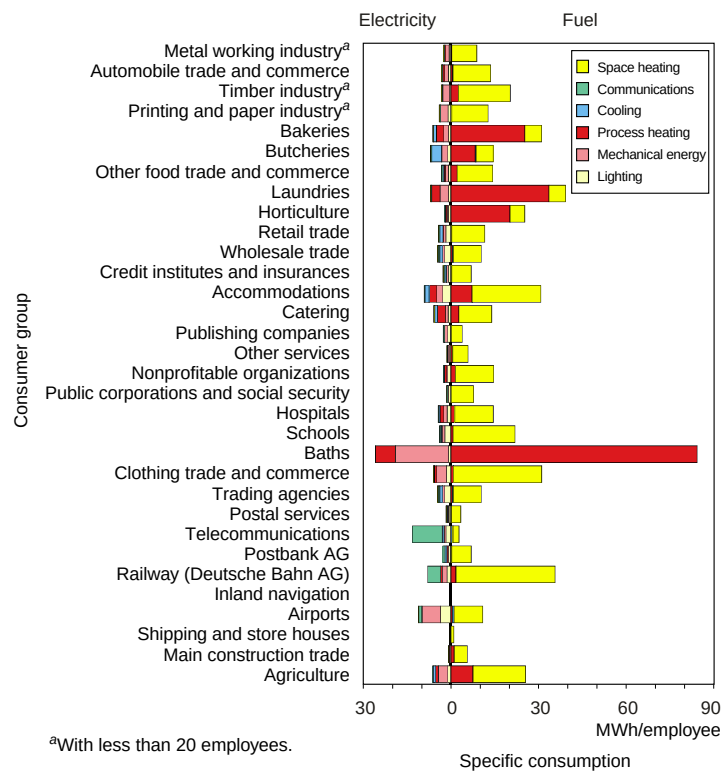


FIGURE 15 Electricity and fuel consumption related to employees; Germany, 1994.

information on the number of employees for each examined branch and the operating area is provided. This background knowledge suggests what sort of energy is consumed and where it is used.

The results for fuels (heating and transportation) and electricity are presented in Fig. 14. Table III and Fig. 14 contain all of the consumer groups, which were classified as small consumer in former times and which are now to a large extent accounted for in the commerce/services sector. These consumer groups also cover agriculture and main construction trade, which are often separated in the international consumer statistics or declared as industry.

The specifications concerning final energy consumption in Fig. 14 (electricity, fuels, space heating, and transportation fuels in the sector commerce/services) represent components of the German energy balance for the commerce/services sector. The given consump-

tion of transportation fuels appears within the transportation sector in the German energy balance.

Figure 14 presents an overview of where the biggest proportions of energy consumption (without transportation fuels) in Germany are situated. The most significant branches are agriculture, retail trade, public corporations and social security, schools, nonprofit organizations, and other services. These alone are responsible for about 50% of the final energy consumption in the commerce/services sector.

Figure 15 presents information on the specific electricity and fuel consumption, differentiated by purpose of use and based on the number of employees. It is obvious that the specific consumption within in the different consumer groups is very heterogeneous.

The main proportions constitute energy consumption due to baths, laundries, accommodations, bakeries, the clothing trade, and commerce. With regard

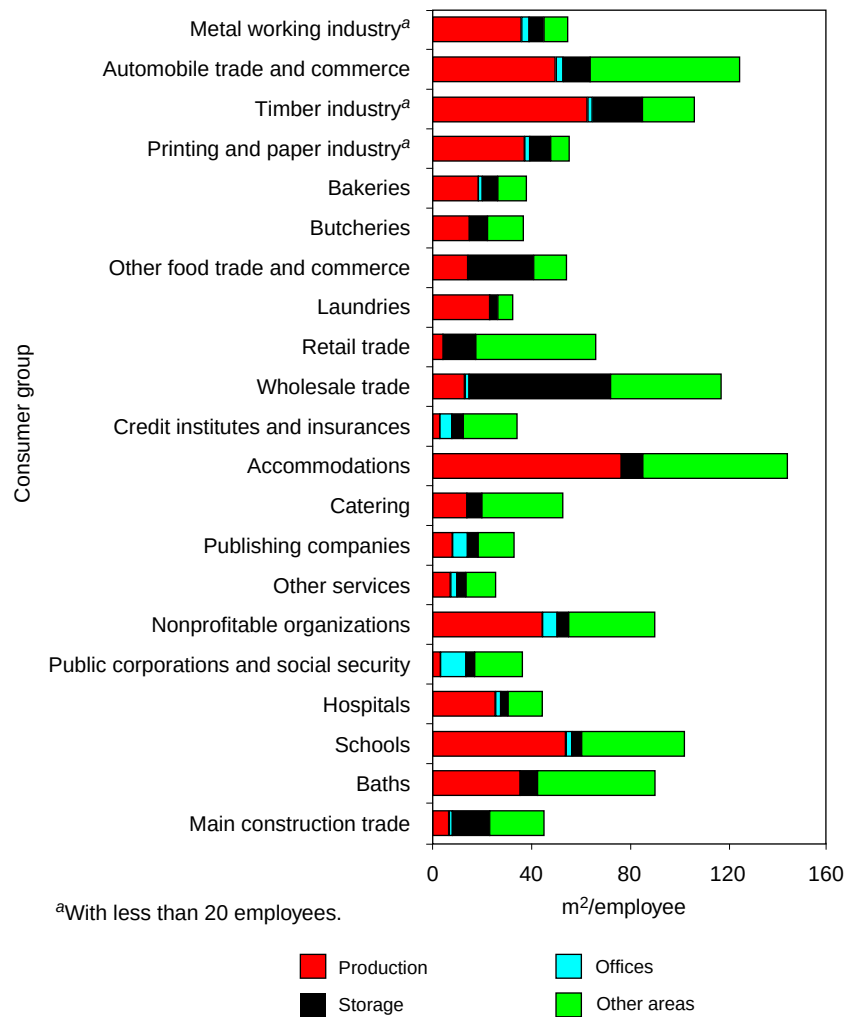


FIGURE 16 Operating area and types of use in commerce/services; Germany, 1994.

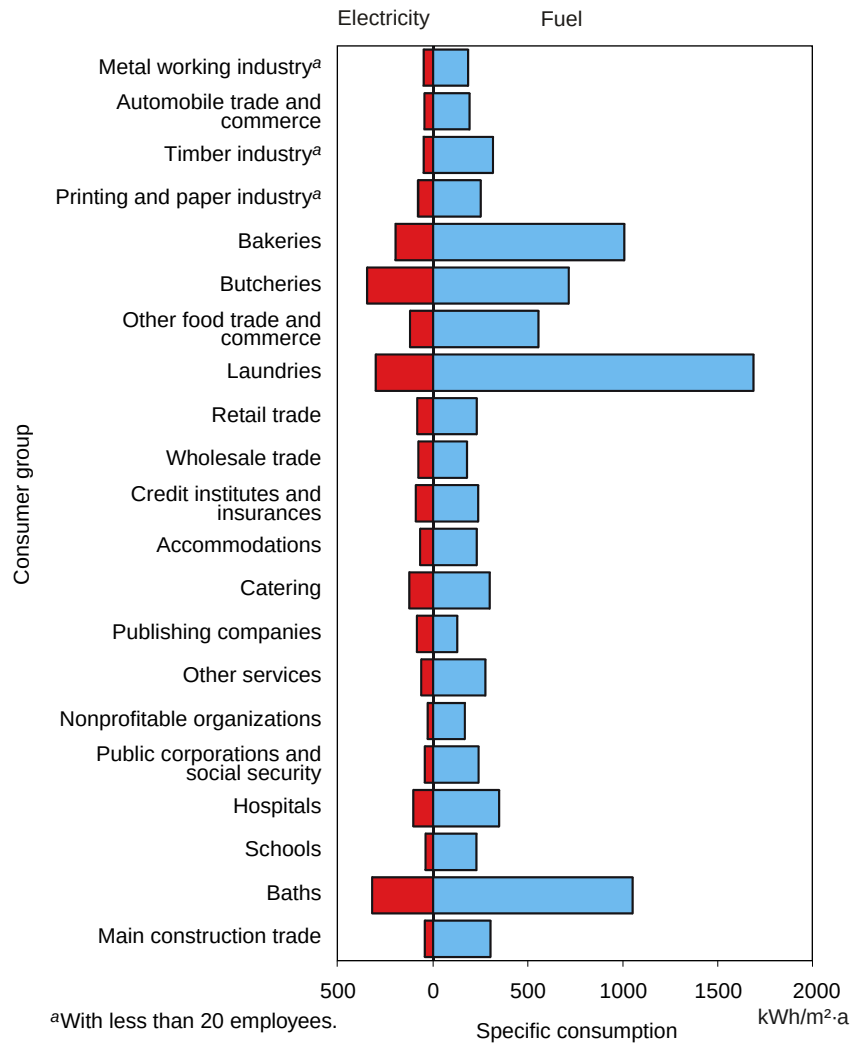


FIGURE 17 Electricity and fuel consumption related to operating area; Germany, 1994.

to the purpose of energy consumption, a differentiation is carried out by space heating, communication, cooling, process heat, mechanical energy, and lighting. Thus, the average proportions of energy consumption for certain applications are known for each consumer group. Note that the operationally effective area has a major influence on the space heating energy consumption. Production and office areas are always heated; warehouse and remaining areas are partly heated or kept at a moderate temperature.

The operating area per employee ranges from 30 m²/employee (other services) to 150 m²/employee (accommodation). By applying this information, the electricity and fuel consumption can be quoted by square meter operating area (see Fig. 16).

Figure 17 suggests that many branches with a very low or negligible process heat demand have a similar

level of electricity and fuel consumption. Seventy-five percent of the consumer groups show a specific consumption in the range of 150 kWh/m² to 250 kWh/m², suggesting that these branches have a demand for space heating and hot water only.

Bakeries, butcheries, other food trade, and commerce, laundries, hospitals, and baths have a disproportionately high demand for fuels that process heat. Beyond this, these consumer groups are characterized by an extremely high specific electricity demand.

5. SUMMARY

This article explores the energy consumption of the commerce/services sector worldwide and particularly

in Germany. When comparing cross-national data, it has to be taken into account, that an accurate demarcation of energy consumption sectors is not always given. By analyzing surveys and data collected onsite, a significant analysis of the electricity and fuel consumption of the commerce/services sector is possible, as the example shows for Germany.

Energy consumption in the commerce/services sector mainly depends on economic prosperity, climate, and the technological status of development. One major influence on energy consumption is the conditioning of workspaces. Lower temperatures and larger workspaces per jobholder—as in the northern industrialized countries—cause higher fuel amounts of consumption for space heating, just as air conditioning in hot climates causes a higher demand for electricity. With economical development, technical standards and instrumentation improve and cause rising electricity consumption.

The analyses presented here are preconditions for better individual consumption consultation, strategic programs toward energy conservation, and an emphasis on technical development.

SEE ALSO THE FOLLOWING ARTICLES

Commercial Sector and Energy Use • Consumption, Energy, and the Environment • Corporate Environmental Strategy • Energy Services Industry • Global Energy Use: Status and Trends • Industrial Energy Efficiency • Industrial Energy Use, Status and Trends • Industrial Symbiosis • International Comparisons of Energy End Use: Benefits and Risks

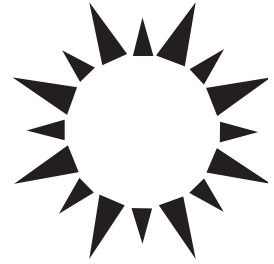
Further Reading

- Geiger, B., Gruber, E., and Megele, W. (1999). Energieverbrauch und Energieeinsparung. In "Handel, Gewerbe, und Dienstleistung." Physica-Verlag.
- Geiger, B., Harrison, A. (2000). "The State of Knowledge Leaders." IEA Workshop on Indicators of Commercial/Service Sector Activity, Energy and CO₂ Emissions, and Policy Impacts, Karlsruhe.
- World Energy Council. (1998). International Energy Data 1998. World Energy Council, London.

Sociopolitical Collapse, Energy and

JOSEPH A. TAINTER

Rocky Mountain Research Station
Albuquerque, New Mexico



-
1. Nature of Collapse
 2. Energy Basis of Ancient Societies
 3. Theories of Energy Flow and Collapse
 4. Examples of Energy Flow in Sociopolitical Collapse
 5. Energy Flow, Collapse, and Resiliency
 6. Energy Flow in the Aftermath of Collapse
 7. Synthesis and Implications

Glossary

collapse Rapid loss of an established level of social, political, or economic complexity.

complexity A measure of the differentiation in structure and organization of a system.

marginal return In an economic process, the amount of extra output per additional unit of input. Marginal returns may increase, remain constant, or diminish.

problem solving The process of responding to the challenges that confront a society or institution.

resources The inputs of matter, energy, or information that support continuity and problem solving.

sustainability Maintaining, or fostering the development of, the systemic contexts that produce the goods, services, and amenities that people need or value, at an acceptable cost, for as long as they are needed or valued.

Collapse is the rapid simplification of a society. It is the sudden, pronounced loss of an established level of social, political, or economic complexity. Widely known examples include the collapses of Mesopotamia's Third Dynasty of Ur (ca. 2100–2000 BC), Mycenaean society of Greece (ca. 1650–1050 BC), the Western Roman Empire (last emperor deposed 476 AD), and Maya civilization of the lowlands of Guatemala (ca. 250–800 AD). There are at least two dozen cases of collapse that are known from history,

archaeology, or both. States and empires are not the only types of institutions that may rapidly simplify. The entire spectrum of societies, from simple foragers to extensive empires, yields examples of collapse. Since all but a few human societies existed before the development of writing, there may be dozens or even hundreds of cases that are not yet recognized archaeologically. Collapse is therefore a recurrent process, and perhaps no society is invulnerable to it.

1. NATURE OF COLLAPSE

Many scholars have argued that collapse develops from adverse changes in the energy flows on which a society depends. In this school, collapse is thought to follow from such misfortunes as pronounced droughts, depletion of resources, or natural cataclysms. Others have searched for causes internal to societies: poor leadership, mismanagement, peasant rebellion, or even intangible factors such as changes in morality. A few scholars argue that collapses occur because of chance concatenations of misfortunes and cannot be understood as a general process. Despite this diversity of views, collapses inevitably alter energy flows. Postcollapse societies are by definition simpler, so much so that colloquially they are sometimes referred to as “dark ages.” The simpler societies that follow a collapse require less energy per capita. Collapses often involve massive loss of population so that human energy extraction is greatly reduced within the region affected.

Collapse affects societies proportionate to their level of complexity. Many scholars identify collapses by the loss of such cultural elements as great traditions of art, architecture, and literature (with the decline in the last leading to the term dark age).

Recognizing that collapse can affect societies organized with any degree of complexity, including simple foragers, a conception of collapse based on great cultural traditions is clearly too restrictive. It would not apply, for example, to the collapse of Chacoan society of the U.S. Southwest (ca. 725–1150 AD). Although it is known for remarkable art and architecture, Chacoan society produced no literature.

A more encompassing approach focuses on complexity. Complexity refers to differentiation in structure and to degree of organization. Societies grow more complex as they develop more parts (institutions, levels of hierarchy, and social or occupational roles) and as they become increasingly organized (i.e., as they increasingly constrain the behavior of a society's parts). One of history's clearest trends is that human societies during the past 12,000 years have shown a seemingly inexorable tendency to increase in complexity. This process still continues, more rapidly than ever.

Collapse is the rapid, substantial loss of an established degree of social, political, or economic complexity. Collapses contravene the trend of social history toward greater complexity. In a society that has collapsed there are fewer institutions, occupations, social roles, and hierarchical levels. There is less flow of information and a lower capacity for organization. People know less about the world beyond their own community. The scale of society shrinks. The Roman Empire, for example, left in its aftermath a multitude of smaller polities, none of which could match Rome's capabilities in organization, engineering, energy extraction, or scale of information flow.

2. ENERGY BASIS OF ANCIENT SOCIETIES

Before the development of fossil fuels, societies were powered largely by solar energy. Through photosynthesis, solar energy produced fuels such as wood and the agricultural produce on which all else depended. Solar energy drove the weather patterns to which forests and farms were adapted. These weather patterns made sailing possible and thus overseas commerce. The products of photosynthesis enabled people to have animals for farm, transport, and military use and animal products on one's table. Photosynthesis supported peasant populations that provided food for the cities, sons for the army, and taxes for the state. The combination of peasant

population, weather patterns, and favorable terrain made irrigation possible in some places. A society's natural endowment of raw materials could be put to use only with solar energy. Thus, photosynthesis, transformed into slaves and charcoal, made it possible to mine and smelt metal-bearing ores. Metals were in turn used to obtain more energy or products of energy. Works of art, architecture, and literature were possible in large part because of the solar energy embedded in artists, architects, and writers; in the production of the raw materials in which artists and architects worked; and in the precious metals with which they were remunerated. Ancient states and empires relied fundamentally on agricultural taxes. Although we marvel today at the monuments and artistic works of ancient societies, these were never more than a small part of ancient economies. A. H. M. Jones, the noted historian of the late Roman Empire, estimated that the Roman government's income was based 90% on agricultural taxes. Trade and industry were comparatively undeveloped.

The persistent weakness of ancient states and empires was that the peasant producers, whose taxes funded government and the military, lived on very small margins of production. Subsistence farmers generate little surplus per capita and often none at all. Taxes were a way to compel peasants to produce more, but excessive taxation could undermine peasants' well-being and ultimately the state. Photosynthesis produces little energy per unit of land (maximum of approximately 170 kcal/m²/year), and it takes much labor and other inputs to appropriate increasingly more of that for human use.

In ancient societies in general, agriculture could support only approximately one person per arable hectare. Vaclav Smil estimates that in Roman or early medieval times, wheat farming could yield approximately a 40-fold net energy gain. This is before subtracting seed requirements and storage losses. (The figure for 1900 AD is approximately 500-fold.) In many areas, this harvest was gained by supplementing human labor with draft animals (which require farmland and labor) and with extensive manuring. Between 180 and 250 h of human labor were combined with approximately 200 h of animal labor to produce approximately one-half metric ton of grain per hectare. In the Middle Ages, European farmers had to set aside one-third to one-half of a crop for the next year's seed. Roman-era farmers in Europe would have faced the same constraint. As discussed later, a late Roman peasant owed between one-fourth and one-third of the gross yield in taxes.

After adding seed needed for sowing to that sold for taxes, it is clear that Roman farmers and their families subsisted on but a small fraction of what they produced. Tax rates were fixed, so when a harvest was low, as frequently it would be, the margin available to support the farmer's family could be inadequate. Tax rates could harm an ancient society by making it difficult for peasants to raise families large enough to sustain the population. The consequence would be a decline in future taxes and in military manpower.

This was the characteristic pattern in ancient states and empires, which were funded by aggregating the small surpluses of many thousands to millions of impoverished subsistence farmers. Thus, the dilemma faced by ancient states was that solar energy severely limited the revenues that they could raise. Often, the direction that ancient rulers chose was to raise revenues in the short term, not realizing that this would undermine the capacity of the productive system to generate future revenue. When confronted with fiscal limits, ancient governments were perpetually tempted to expand spatially, thereby capturing more of the earth's surface where solar energy falls.

3. THEORIES OF ENERGY FLOW AND COLLAPSE

There are many theories of why societies collapse. As emphasized in the following discussion, however, most of these fail in one way or another to explain collapse fully. After describing these theories and their limitations, a newer approach is introduced.

3.1 Approaches to Energy Flow and Collapse

Students of collapse frequently blame interruptions in the resources on which a society depends. Without sufficient energy flow to sustain people and their institutions, collapse must follow. This may occur either because a resource is used to depletion or because external events, such as a change in climate, cause a once-dependable resource to become insufficient.

Such theories have been developed for a number of collapses. J. Donald Hughes, for example, argued that the ultimate cause of Rome's collapse was environmental deterioration, especially deforestation. The American geographer Ellsworth Huntington believed that climate change caused the barbarian migrations that the Roman Empire could

not survive. Karl Butzer, a noted scholar of paleoenvironments, argued that insufficient Nile floods caused the collapse of Egypt's Old Kingdom (ca. 3100–2181 BC) and exacerbated other political upheavals in Egyptian history. The spectacular collapse of the Southern Lowland Classic Maya has prompted scholars to search for environmental factors, such as erosion, soil nutrient depletion, or conversion of forests to savanna grasslands. Rhys Carpenter, an accomplished student of Mycenaean civilization, argued that the Mycenaean collapse came from drought, which caused famine, depopulation, and migration. Barbara Bell argued that two so-called dark ages encompassing the eastern Mediterranean and Near East, ca. 2200–2000 and 1200–900 BC, can both be explained by widespread droughts that lasted several centuries.

The arid lands of Mesopotamia (modern Iraq, eastern Syria, and southeastern Turkey) provide some of the most compelling examples of resource depletion and collapse. There were two catastrophic collapses here: those of the Third Dynasty of Ur (ca. 2100–2000 BC) and the Abbasid Caliphate (749–1258 AD). In each case, the collapse has been explained as resulting from overirrigation of desert soils, which caused soil salinization that destroyed agricultural productivity. Sociopolitical collapse was accompanied by depopulation and environmental damage that in each case lasted centuries. Working in upper Mesopotamia at the site of Tell Leilan, Syria, Harvey Weiss found evidence that led him to suggest that collapse ca. 2200 BC in this region was caused by an abrupt change in climate that affected hundreds of thousands of both agricultural and pastoral peoples.

Some scholars focus more on human than natural resources to understand collapse. Students of the Maya periodically reformulate the idea that the stresses of supporting the Maya ruling classes, which involved mandatory labor to build lavish limestone monuments, sparked a peasant revolt that toppled the elites and their cities. Similar ideas have been advanced for collapses in Mesopotamia, Highland Mesoamerica, Peru, China, and elsewhere. In regard to the Roman Empire, Boak and Sinnigen argued that "Rome failed to develop an economic system that could give to the working classes of the Empire living conditions sufficiently advantageous to encourage them to support it devotedly and to reproduce in adequate numbers." Even in the fourth-century AD, Ammianus Marcellinus attributed the problems of the later Roman Empire to the growth of bureaucracy and excessive taxation.

A few historians, archaeologists, and other writers indict unpredictable natural cataclysms in collapses. These alter energy flows so wrenchingly that continuity or recovery are impossible. Hurricanes, earthquakes, and epidemics have all been implicated in the Maya collapse. As early as 1939, Marinatos suggested that the collapse of Minoan society on Crete (ca. 2000–1200 BC) was caused by the great eruption of the Aegean island of Thera. Popular attempts to link Plato's parable of Atlantis to an actual place have kept this idea alive. Recent writers have elaborated on the theme, even extending the effects of the devastation to the entire eastern Mediterranean.

3.2 Assessment of Approaches to Energy Flow and Collapse

Vaclav Smil observed that many collapses occurred without any evidence of weakened energy bases. Thus, energy flow alone cannot explain collapse as a recurrent historical pattern. Every society has arrangements to ameliorate changes in energy flow. Indeed, societies have withstood even pronounced constrictions in energy flow, provided that they prove temporary. Nineteenth-century examples of pronounced disruption in energy flow include the Irish potato famine and the great eruption of Krakatoa in the South Pacific. These are sometimes compared to the disasters of antiquity (e.g., the eruption of Thera), but neither caused a collapse. Catastrophes fail to explain collapse because societies withstand disasters all the time without collapsing. Moreover, sometimes the date of the catastrophe fails to match the date of the collapse.

Changes in climate also seem inadequate to account for the matter. In studies of climate and collapse there are problems of logic. In the cases of Tell Leilan, Syria, and the Hohokam of the U.S. Southwest, archaeologists have postulated that deterioration in energy flows caused by changing climate caused, at different times, both collapse of complexity and its increase. These scholars do not explain how climate change can have such opposite effects, leaving an apparent contradiction in their arguments. In the case of drought and the Mycenaean collapse, the part of Greece with the least rainfall, Attica, was the last to collapse. If deforestation contributed to the Roman collapse, why did the same problem, later in European history, help spur economic development, as Wilkinson has shown? A change in resources can by itself rarely explain collapse.

There is no doubt that energy flow is almost always involved in collapse. Often, energy flow

changes in the aftermath of simplification and depopulation. Rarely, if ever, has a change in energy flow been the sole cause of a collapse. To understand the role of energy in collapse, it must be considered in the context of social, political, and economic changes.

3.3 Embedding Energy Flow and Collapse within Economics and Problem Solving

A new approach embeds energy in the social, political, and economic changes that a society undertakes to solve problems. Problem solving among humans (and even among animals) shows a general tendency to move from simple, general, and inexpensive (known colloquially as plucking the lowest fruit first) to complex, specialized, and costly. Sustainability or collapse follow from the long-term success or failure of problem-solving institutions. Throughout human history, a primary problem-solving strategy has been to increase complexity. Often, we respond to problems by creating new and more sophisticated technologies, new institutions, new social roles, or new hierarchical levels or by processing more information. Each of these strategies increases the complexity of a society. Problem solving has been much of the force behind the increasing complexity of human societies, especially during the past 12,000 years.

Changes in complexity can often be enacted quickly, which is one reason why we resort to complexity to solve problems. This is illustrated by the response to the attacks on the United States on September 11, 2001. Much of the immediate response to the problem of preventing future attacks involved increasing the complexity of public institutions by establishing new agencies, absorbing existing ones into the federal government, and expanding the scale and degree of control over realms of behavior from which a threat might arise. Such changes can be implemented more rapidly than can changes in military technology, in the technology and other techniques of espionage, or in the technology to detect explosives in air travel baggage.

As a problem-solving strategy, complexity often succeeds. The difficulty is that complexity costs. In any living system each increase in complexity has a cost, which may be measured in energy or other resources or in transformations of these, such as money, labor, or time. Before the development of fossil fuels, increasing the complexity of a society typically meant that people worked harder. This is still often the case. In the response to the September 11, 2001, attacks, anyone who travels by air understands that doing so has become more complex, and

the personal cost has risen. The cost is reflected not just in the price of travel (which is controlled by competition) but also in such intangibles as time spent standing in lines, the indignity of personal searches, and greater care regarding what personal items one travels with. These are examples of the subtle costs of complexity in problem solving.

If an increase in complexity resolves a problem, it is likely to be adopted for its short-term value. The costs of complexity are cumulative and long term. As simple, inexpensive solutions are adopted for problems in such areas as organization, information flow, or technology, the strategies that remain to be deployed are increasingly complex and costly. Complexity is an economic function and can be evaluated by returns on effort. Early efforts at problem solving often produce simple, inexpensive solutions that may yield positive returns. Yet as the highest-return ways to produce resources, process information, and organize a society are progressively implemented, problems must be addressed by more costly and less effective responses. As the costs of solutions increase, the marginal return to complexity in problem solving begins to decline. Carried far enough, diminishing returns to problem solving cause economic stagnation and disaffection of the support population. As illustrated later, a prolonged period of diminishing returns to complexity in problem solving makes a society vulnerable to collapse.

4. EXAMPLES OF ENERGY FLOW IN SOCIOPOLITICAL COLLAPSE

Energy, leadership, political organization, taxation, and competition are all involved in collapse. To illustrate this, I describe five cases in detail, including two collapses in Mesopotamia in which energy and collapse are clearly linked, although not as simply as sometimes described. The case of the Lowland Classic Maya exemplifies how competition and pressures to produce great monuments create stress on peasantry that undermines the production system. The Roman and Byzantine empires illustrate how complexity in problem solving makes a society vulnerable to collapse. In the Roman case, the collapse lasted centuries. In the Byzantine example, the collapse actually provided a basis for recovery.

4.1 Mesopotamia

Ancient Mesopotamia produced two cases that illustrate the relationship of collapse to energy—the

collapses of the Third Dynasty of Ur (ca. 2100–2000 BC) and the Abbasid Caliphate (749–1258 AD).

In southern Mesopotamia, intensive irrigation initially increases agricultural yields that support growing prosperity, security, and stability. The Third Dynasty of Ur pursued this course. It expanded the irrigation system and encouraged growth of population and settlement. To capture this enhanced energy flow, it established a vast bureaucracy to collect taxes. This strategy plants the seed of its own demise, however. After a few years of overirrigating, saline groundwaters rise and destroy the basis of agricultural productivity. The political system loses its resource base and is destabilized. Large irrigation systems that require central management are useless once the state lacks the resources to maintain them.

A few centuries earlier, in the Early Dynastic period (ca. 2900–2300 BC), crop yields averaged approximately 2030 liters per hectare. Under the Third Dynasty of Ur this declined to 1134 liters. At the same time, Third Dynasty of Ur farmers had to plant their fields at an average rate of 55.1 liters per hectare, more than double the previous rate. Badly salinized lands go out of production almost indefinitely, so the pressure intensifies to get the most from the remaining fields. As yields declined and costs rose, farmers had to support an elaborate state structure. It was a system that took a high toll in labor. The Third Dynasty of Ur was following a strategy of costly intensification that clearly yielded diminishing returns.

The Third Dynasty of Ur hung on through five kings and then collapsed, with catastrophic consequences for the larger population. By 1700 BC, yields were down to approximately 718 liters per hectare. Of the fields still in production, more than one-fourth yielded on average only approximately 370 liters per hectare. The labor demands of farming a hectare of land were inelastic, so for equal efforts cultivators took in harvests approximately one-third the size of those a millennium earlier. Soon the region was extensively abandoned. By a millennium or so after the Third Dynasty of Ur, the number of settlements had decreased 40% and the settled area contracted by 77%. Population densities did not rebound to those of the Third Dynasty of Ur until the first few centuries AD, a hiatus of 2500 years.

Rulers of Mesopotamia, from the last few centuries BC into the Islamic period, repeated the strategy tried 2000 years earlier by the Third Dynasty of Ur. Population increased fivefold in the last few centuries BC and the first few centuries AD. The number of urban sites grew 900%. These trends

continued through the Sassanian period (226–642 AD), when population densities came to exceed significantly those of the Third Dynasty of Ur. At its height, the area of settlement was 61% greater than during the Third Dynasty of Ur.

Under Khosrau I (AD 531–579) the Sassanian dynasty reached its height, but the needs of the state took precedence over peasants' ability to pay. Taxes were no longer remitted for crop failure. Because the tax was fixed whatever the yield, peasants were forced to cultivate intensively. State income rose sharply under Khosrau II (590–628). This level of income would have been needed for the perpetual wars in which he engaged with the Byzantine Empire. Under the Abbasid Caliphate, an Islamic dynasty, tax assessments increased in every category. Fifty percent of a harvest was owed under the Caliph Mahdi (775–785), with many supplemental payments. Sometimes, taxes were demanded before a harvest, even before the next year's harvest.

Under the Abbasids there was unprecedented urban growth. Baghdad grew to five times the size of 10th-century Constantinople and 13 times the size of the Sassanian capital, Ctesiphon. The capital was moved often and each time built anew on a gigantic scale. The Caliph al-Mutasim (833–842) built a new capital at Samarra, 120 km upstream from Baghdad. In 46 years, he and his successors built a city that stretched along the Tigris for 35 km. It would have dwarfed imperial Rome.

As the state undertook these constructions, however, it did not always fulfill its irrigation responsibilities. As the irrigation system grew in size and complexity, maintenance that had once been within the capacity of local communities was no longer so. Communities came to depend on the imperial superstructure, which in turn became increasingly unstable. Peasants had no margins of reserve, and revolts were inevitable. Civil war and rebellion meant that the hierarchy could not manage the irrigation system. Mesopotamia experienced an unprecedented collapse. In the period from 788 to 915 revenues fell 55%. The Sawad region, at the center of the empire, had supplied 50% of the government's revenues. This dropped within a few decades to 10%. Most of this loss occurred between the years 845 and 915. In many once-prosperous areas, revenue declined 90% within a lifetime. The perimeter of state control drew inward, which diminished any chance to resolve the agricultural problems. By the early 10th century irrigation weirs were listed only in the vicinity of Baghdad. In portions of Mesopotamia the occupied area shrank

by 94% by the 11th century. The population declined to the lowest level in five millennia. Urban life in 10,000 square kilometers of the Mesopotamian heartland was eliminated for centuries.

Collapses in Mesopotamia present clear evidence of the role of energy flow disruptions. When the resource base was destroyed by excessive irrigation, population levels and political complexity could not be maintained. Yet the clarity of these cases is deceptive because it obscures a fundamental question: Why did Mesopotamian rulers push their support systems to the point at which large populations and their own political systems were endangered? Such a question cannot be answered here. Rather, it underscores the point made previously that energy flows, although vital in understanding collapse, do not fully explain it. This will be clear in the next two cases.

4.2 The Southern Lowland Classic Maya

The Southern Lowland Classic Maya (ca. 250–800 AD) present one of history's most enigmatic collapses. The Maya Lowlands (Fig. 1) were one of the most densely populated areas of the preindustrial world, with an average of 200 persons per square kilometer. Tikal, one of the major Maya cities, had a population estimated at nearly 50,000, which is roughly the same as ancient Sumerian cities. In the southern part of the area known today as Quintana Roo, there are stretches of terrain 40–50 km long in which there is no gap of more than 100 m between Maya structures. The Maya are best known for their art, their cities and temples, and, as Robert Netting phrased it, their "abrupt departure from the stage of world history."

To support so many people the southern Lowland Maya transformed their landscape. Over much of the area the natural rainforest was cleared. One technique they employed was to channel and store water. Although the area today is a high tropical forest, precipitation varies seasonally, and many areas have little surface water. The Maya accordingly built canals, dams, reservoirs, and small wells and in Yucatán modified the large limestone sinkholes known as *cenotes*. There are 180 km of canals along the Rio Candelaria. They represent at least 500,000 days of labor to move (entirely with human workers) 10 million cubic meters of earth.

Hillsides across the Lowlands were often terraced for planting. Terracing directs water flow, traps silt, and creates cultivable land. Terraces are spread across more than 11,000 square kilometers. Much of the southern Lowlands (the area known as the



FIGURE 1 The Maya Lowlands and major archaeological sites.

Petén) also contains walled fields, a type of feature that indicates intensive, sustained cultivation.

Perhaps the most interesting agricultural features of the Lowlands are systems of raised fields and associated canals. These were built to supply dry, cultivable land in areas otherwise inundated. Raised fields were built within lakes and lagoons and at river margins, but they were primarily built within swamps, thereby reclaiming land that could be highly productive. Modified and intensively culti-

vated swamps may have become the most valuable land. Not surprisingly, the larger cities tend to be located near swamps. By building raised fields in such areas, the Maya could have maintained a moist, root-level environment; fertilized fields with rich, organic muck from canal bottoms; and perhaps raised fish as well. Tended carefully, such fields could have supported continuous cropping.

By the use of such facilities the energy of the Lowlands was channeled to produce a simplified

biota consisting of maize, squash, avocado, cacao, cotton, root crops, and fruit-bearing trees. Simplified food webs became the foundation for a diverse and complex society. Ranks of bureaucrats, artisans, elites, and paramount rulers held sway in cities whose centers were built of gleaming limestone. Their sense of time was based on astronomical observations so precise that they were not equaled in Europe or China for centuries to come. The Maya developed a system of writing that has only recently been deciphered. Their art is greatly admired today. All of this complexity was based on the energy flow made available from their managed landscape.

One consequence of the development of complex societies in the Maya area is that the cities were often at war. Warfare is a prominent theme in Maya art, with many scenes of domination and execution of captives. There are great earthen fortifications at such sites as Becan, Calakmul, Oxpeul, and Tikal. The defensive works at Tikal include a moat and earthwork 9.5 km long. Warfare in turn reinforced centralized political authority.

Maya complexity imposed a high cost on the support population. Great efforts were required to transform the forest, keep the land from reverting to forest, and keep the soil fertile; to support non-producing classes of bureaucrats, priests, artisans, and rulers; and to build and maintain the great cities. Over time these costs actually grew. The construction of monuments increased over the years. Across the southern Lowlands, 60% of all dated monuments were built in a period of 69 years, between 687 and 756 AD.

All this effort took its toll on the support population. At Tikal, by approximately 1 AD, the elites began to grow taller than the peasants, and within a few centuries the difference averaged 7 cm. That elites should have enjoyed better nutrition during childhood is not surprising. By the late Classic (ca. 600–800 AD), however, the stature of adult males actually dropped among both elites and peasants. The nutritional value of the Tikal diet had apparently declined for both rulers and ruled. At the site of Altar de Sacrificios there was a high incidence of malnutrition and/or parasitic infections. Childhood growth was interrupted by famine or disease. Many people suffered from anemia, pathologic lesions, and scurvy. At Altar de Sacrificios, the stature of adult males declined in the first few centuries AD, and life expectancy declined in the late Classic. Not surprisingly, the population of the southern Lowlands, which had grown during the first several centuries AD, leveled off between 600

and 830 and thereafter grew no more. Such population trends often foreshadow collapse.

The magnificence of late Maya cities clearly came at the expense of the peasant population, and ultimately at the cost of the entire system's sustainability. It was a classic case of diminishing returns to complexity. As the complexity of the political system and the cities increased through time, the costs of the system, in human labor, also increased. It is clear that the marginal return to these investments was declining by the late Classic. Ever greater investments in military preparedness, monumental construction, and agricultural intensification brought no proportionate return in the well-being of most of the population. To the contrary, as the demands on the peasant population increased, the benefits to that population actually declined. This was particularly the case in the seventh and eighth centuries, when a great increase in laborious monumental construction fell on a stressed and weakened population that was no longer growing.

The consequences were predictable. One by one, the Maya polities began to collapse and their cities were progressively abandoned (Fig. 2). At the same time, the southern Lowlands experienced one of history's major demographic disasters. In the course of approximately a century, the population declined by 1–2.5 million people and ultimately declined to a fraction of its peak. There were not enough Maya left to control the landscape, nor did those remaining need to farm as they once had. Intensive agriculture was abandoned and the forests regrew. Cultural complexity in this region was based on an engineered landscape with low species diversity, short food chains, and diversion of energy from plant to human

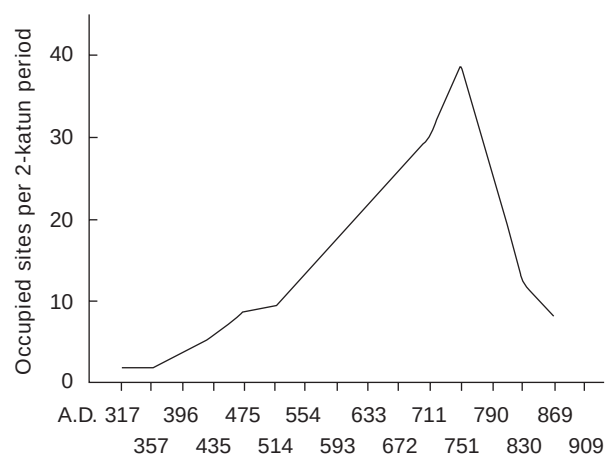


FIGURE 2 Occupation of Classic Maya centers. One katun is approximately 20 years.

biomass. In the long term, although the Lowlands landscape could seemingly support dense human populations, it could not sustain the complex political systems that large populations often require.

4.3 Collapse of the Western Roman Empire

The Western Roman Empire (Fig. 3) is one of history's most instructive collapses. The Romans' early success came from a means of expansion that was fiscally self-perpetuating. Defeated peoples gave the economic basis, and some of the manpower, for further expansion. It was a strategy with high economic returns. By 167 BC, for example, the Romans were able to eliminate taxation of themselves and still expand the empire. Yet the economics of an empire such as the Romans assembled are deceptive. The returns to any campaign of conquest are highest initially, when the accumulated surpluses of the conquered peoples are appropriated. Thereafter, the conqueror assumes the cost of administer-

ing and defending the province. These responsibilities may last centuries and are paid from yearly agricultural surpluses. The fiscal basis of such an empire shifts from accumulated surpluses (generally embodied in precious metals, works of art, and population) to current solar energy.

Once the phase of conquest was complete, and the accumulated surpluses spent, the Roman government was financed by agricultural taxes that barely sufficed for ordinary administration. When extraordinary expenses arose, typically during wars, the precious metals on hand frequently were insufficient to produce the required coinage. Facing the costs of war with Parthia and rebuilding Rome after the Great Fire, Nero began in 64 AD a policy that later emperors found irresistible. He debased the primary silver coin, the denarius, reducing the alloy from 98 to 93% silver. It was the first step down a slope that resulted two centuries later in a currency that was worthless (Fig. 4) and a government that was insolvent.

In the half century from 235 to 284, the empire nearly came to an end. There were foreign and civil



FIGURE 3 The Roman Empire in 117 AD.

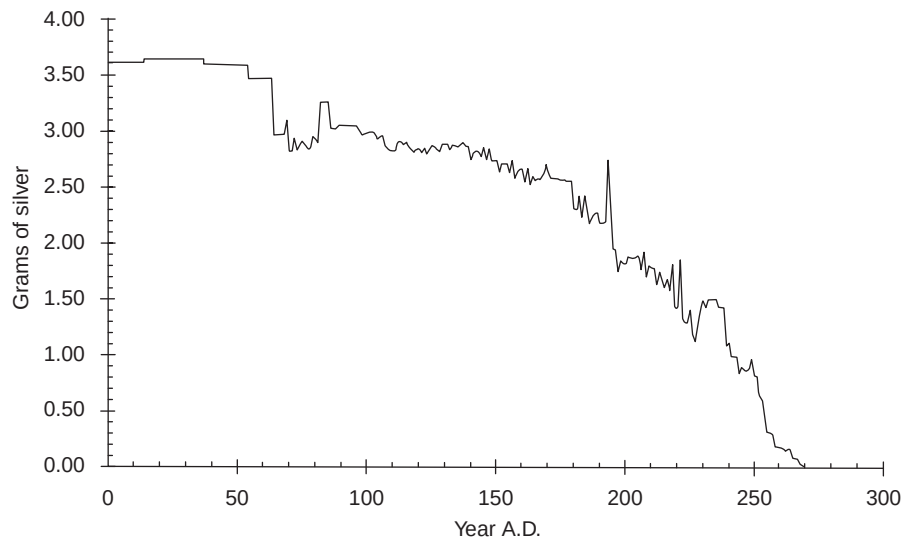


FIGURE 4 Debasement of the Roman silver currency to 269 AD.

wars almost without interruption. The period witnessed 26 legitimate emperors and perhaps 50 usurpers. Cities were sacked and frontier provinces devastated. The empire shrank in the 260s to Italy, the Balkans, and North Africa. By prodigious effort the empire survived the crisis, but it emerged at the turn of the fourth-century AD as a very different organization.

In the late third and early fourth centuries, Diocletian (284–305) and Constantine (306–337) designed a government that was larger, more complex, more highly organized, and much more costly. They doubled the size of the army, always the major part of imperial costs. To pay for this, the government taxed its citizens more heavily, conscripted their labor, and dictated their occupations. Increasingly more of the empire's energy was devoted to maintaining the state.

Diocletian established Rome's first budget, and each year a tax rate was calculated to provide the revenue. The tax was established from a master list of the empire's people and lands, tabulated down to individual households and fields. In an era when travel and communication were slow, expensive, and unreliable, it took substantial organization and personnel just to establish and administer a tax system so minutely detailed. Taxes increased, apparently doubling between 324 and 364. Villages were liable for the taxes on their members, and one village could even be held liable for another. Tax obligations were extended to widows, orphans, and dowries. Even still, to meet its needs, the government had to conscript men for the army and requisition services.

Occupations were made hereditary and obligatory. Positions that had once been eagerly sought, such as in city senates, became burdensome because leading citizens were held responsible for tax deficiencies.

The tax system supporting the more complex government and larger army had unforeseen consequences. After plagues decimated the population in the second and third centuries, conditions were never favorable for recovery. There were shortages of labor in agriculture, industry, the military, and the civil service. The tax system of the late empire seems to have been to blame because the rates were so high that peasants could not accumulate reserves nor support large families. Whatever crops were brought in had to be sold for taxes, even if it meant starvation for the farmer and his family. Farmers who could not pay their taxes were jailed, sold their children into slavery, or abandoned their homes and fields. In such circumstances it became unprofitable to cultivate marginal land because too often it would not yield enough for taxes and a surplus. Despite government edicts, marginal lands went out of cultivation. In some provinces, up to one-third to one-half of arable lands came to be deserted by the late empire. Faced with taxes, peasants would abandon their lands and flee to the protection of a wealthy landowner, who was glad to have the extra labor.

On the death of the emperor Theodosius in 395 AD, the Roman Empire was divided into eastern and western parts, each of which had to raise its own revenues and most of its own troops. The east was by far the wealthier half, with fewer problems at its borders. From the late fourth century the barbarians

could no longer be kept out. They forced their way into Roman lands in western Europe and North Africa, initially causing great destruction. The government had no choice but to acknowledge them as legitimate rulers of the territories they occupied. Throughout the fifth century the western empire was in a positive feedback loop tending toward collapse. Lost or devastated provinces meant lower government income and less military strength. Lower military strength in turn meant that more areas would be lost or ravaged.

In the 20 years following the death of emperor Valentinian III (455), the western Roman army disappeared. The government came to rely almost exclusively on mercenaries from Germanic tribes. Finally, these could not be paid. They demanded one-third of the land in Italy in lieu of pay. This being refused, they revolted and deposed the last emperor in Italy, Romulus Augustulus, in 476.

The strategy of the later Roman Empire was to respond to a near-fatal challenge in the third century by increasing the size, complexity, power, and costliness of the primary problem-solving system—the government and its army. The higher costs were undertaken not to expand the empire or to acquire new wealth but to sustain the status quo. The benefit/cost ratio of imperial government declined as it lost both legitimacy and support. In the end, the Western Roman Empire could no longer afford the problem of its own existence.

5. ENERGY FLOW, COLLAPSE, AND RESILIENCY

It is common to think of collapse as a catastrophe, involving the loss of great traditions of art, architecture, and literature. In the case of the early Byzantine Empire (the Eastern Roman Empire, 395–1453 AD), though, the reduction of complexity that is the defining characteristic of collapse contributed to resiliency and recovery. The Byzantine case may be unique among collapses. It is especially instructive in regard to the economics of complexity in problem solving.

The Eastern Roman Empire survived the fifth-century debacle and met its end only when the Turks took Constantinople in 1453. The Byzantine Empire lost territory throughout much of its history and several times nearly fell. Yet it survived near catastrophes through one of history's most unusual strategies.

The eastern emperor Anastasius (491–518) established a sound coinage in the copper denominations

on which daily life depended. He gave the army generous allowances for rations, arms, and food. Commerce was revitalized and the army again attracted native volunteers. Within a few decades these reforms had produced such results that Justinian (527–565) could attempt to recover the western provinces.

An army sent to North Africa in 532 conquered the Kingdom of the Vandals within 1 year. Almost immediately, the Byzantine general Belisarius was sent to reconquer Italy. He had captured the Ostrogothic King and nearly all of Italy when he was recalled in 540 to fight the Sassanian Persians.

In 541, when the job in Italy seemed about done, bubonic plague swept over the empire. It had not been seen before in the Mediterranean. Like any disease introduced to a population with no resistance, the effects were devastating. Just as in the 14th century, the plague of the sixth century killed one-fourth to one-third of the population. The enormous loss of taxpayers caused immediate financial problems. Army pay fell into arrears, and both Italy and North Africa were nearly lost. By debasing the currency and slashing expenditures, the emperor was able to send another army to Italy. Italy was reconquered by 554, but much of it was lost again when the Lombards invaded in 572. Throughout the rest of the sixth century, costly wars continued in Italy, North Africa, the Balkans, and with the Persians.

A rebellion of the Byzantine army in 602 rekindled the Persian war and began a crisis that lasted more than a century and nearly brought the empire to an end. The war against the Persians went badly. Sensing Byzantine weakness, Slavs and Avars overran the Balkans. North Africa and Egypt rebelled and placed Heraclius (610–641) on the Byzantine throne. The empire he took over was financially exhausted. The Persians captured Syria, Palestine, and Egypt, overran Anatolia, and besieged Constantinople from 618 to 626.

In 616, Heraclius cut the pay of troops and officials in half. He followed his predecessors by lowering the weight of the primary copper coin, the *folles* (Fig. 5). Heraclius' economic measures bought time for his military strategy to work. In 627, Heraclius destroyed the Persian army. The Byzantines got all their lost territory returned. The war had lasted 26 years and resulted in no more than restoration of the status quo of a generation earlier.

The empire was exhausted by the struggle. Arab forces, newly converted to Islam, defeated the Byzantine army decisively in 636. Syria and Palestine, which had taken 18 years to recover, were lost

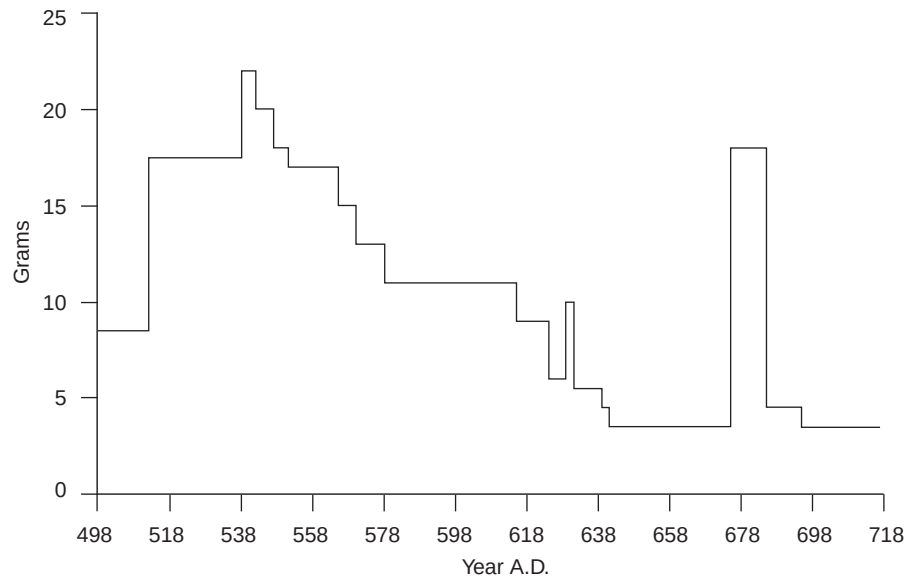


FIGURE 5 Weight of the Byzantine follis, 498–717 AD.

again. Egypt was taken in 641. The wealthiest provinces were permanently gone, and soon the empire was reduced to Anatolia, Armenia, North Africa, Sicily, and parts of Italy. The Arabs conquered the Persian Empire entirely. (In time, parts of the former Persian Empire became the center of the Abbasid Caliphate, described previously.)

Under Constans II (641–668) and throughout the seventh century, the situation continued to deteriorate. It appeared that it would be only a matter of time before the remaining Byzantine lands fell to the Arabs. The Arabs raided Asia Minor nearly every year for two centuries. Constantinople was besieged each year from 674 to 678, then continuously from 717 to 718. The Byzantines defeated both sieges, and this proved to be the turning point in a centuries-long struggle. The Arabs were never again able to mount such a threat.

During this century of crisis the social and economic life of the eastern Mediterranean was thoroughly transformed. Population declined. Monetary standards and economic institutions were destroyed. Around 659, Constans II cut military pay in half again. Army pay now stood at one-fourth its level in 615 so that the economy no longer received large influxes of new coin. By 700 most people of the eastern Mediterranean no longer used coins in everyday life. The economy developed into its medieval form, organized around self-sufficient manors.

Roman emperors of the late third and early fourth centuries had responded to a similar crisis by complexification. They increased the complexity of

administration, the regimentation of the population, and the size of the army. This was paid for by levels of taxation so harmful that lands were abandoned and peasants could not replenish the population. Constans II and his successors could hardly impose more of the same exploitation on the depleted population of the shrunken empire. Instead, they adopted a strategy that is truly rare in the history of complex societies: systematic simplification, in effect a deliberate collapse.

The government had lost so much revenue that even at one-fourth the previous rate it could not pay its troops. Constans devised a way for the army to support itself. Available lands were divided among the troops. Soldiers (and later sailors) were given grants of land on condition of hereditary military service. Constans expected the troops to provide their own livelihood through farming, with a small monetary supplement. Correspondingly, the Byzantine fiscal administration was greatly simplified.

The transformation ramified throughout Byzantine society. Both central and provincial government were simplified. In the provinces, the civil administration was merged into the military. Cities throughout Anatolia contracted to fortified hilltops. Aristocratic life focused on the imperial court. There was little education beyond basic literacy and numeracy, and literature consisted of little more than lives of saints. The period is sometimes called the Byzantine dark age.

The simplification rejuvenated Byzantium. The new farmer-soldiers became producers rather than

consumers of the empire's wealth. From this new class of farmers came the force that sustained the empire. By lowering the cost of military defense the Byzantines secured a better return on their most important investment. The empire began to lose land at a much slower rate. The Arabs continued to raid Anatolia but were unable to hold any of it for long. Soldiers were always nearby. Fighting as they were for their own lands and families, they had much greater incentive and performed better.

Once the threat to the empire's existence was over, campaigns against the Bulgars and Slavs gradually extended the empire in the Balkans. Greece was recaptured. In the 10th century the Byzantines reconquered parts of coastal Syria. Overall after 840 the size of the empire was nearly doubled. The process culminated when Basil II (963–1025) conquered the Bulgars and extended the empire's boundaries again to the Danube. In two centuries the Byzantines went from near disintegration to being the premier power in Europe and the Near East, an accomplishment won by a deliberate collapse in which they systematically simplified, and thereby reduced the costs of, their army, government, and society.

6. ENERGY FLOW IN THE AFTERMATH OF COLLAPSE

Since collapse is the loss of complexity, and higher levels of complexity depend on greater flows of energy, a primary characteristic of postcollapse societies is that they require and produce less energy. This is a consequence of collapse that occurs because there are fewer elite ranks to extract resources from the populace, because expensive sectors of investment (such as the Roman army) disappear, or because a lower degree of social and economic specialization means that peasant households no longer have to support specialists. A late Roman peasant, for example, might have paid one-fourth to one-third of gross yields in taxes if he owned his own land. If he rented land, taxes and rent together took one-half to two-thirds of a crop. Thus, Roman peasants were perpetually impoverished, often undernourished, and sometimes forced to sell children into slavery when they could not be fed. This undermined the productive capacity of the empire and its sustainability. After the plagues of the second and third centuries AD, peasants were unable to replenish the population, and the empire lost much of its energy-producing capacity. In post-Roman

Gaul, in contrast, taxes declined to as low as 10%. Early medieval peasants may have been better off and seen more of their children survive.

The Byzantines responded to lower energy flow by simplifying their social, political, and economic systems and by shortening energy flow networks. In the sixth-century AD, energy would typically have flowed from the sun to farms; from farms to peasants; from peasants to purchasers of grain, thereby converting the grain to coinage; from peasants to tax officials; from tax officials to the government; and from the government to the army. At each step some energy was lost to transaction costs. Just as in a trophic pyramid, energy was lost each time it was transformed or passed to another level. After the mid-seventh century, the energy on which the empire depended flowed through fewer transactions and levels. Energy passed from the sun to farms. From there part of it was harvested and used directly by soldiers, eliminating many intermediate steps. Other peasants still paid taxes, but since the government spent less of this revenue on the army there was more to allocate to other needs. Overall much less energy went to transaction costs, so net energy increased as a proportion of gross production. The empire revived and went on to expand.

In general, postcollapse societies are smaller, with simpler hierarchies and less socioeconomic specialization. They produce less monumental architecture or none at all. Population often declines, sometimes dramatically as in the Maya collapse. Many settlements are abandoned, and those that remain may contract. Long-distance trade declines. The known world shrinks, and over the horizon lies the unknown. There are often declines in literacy and numeracy because the simpler society requires less of these. Since there are fewer elites for whom to produce art, and fewer artisans, postcollapse societies produce less art. All these changes involve reductions in energy extraction and flow, either as cause or as consequence. In some cases, energy flow is so greatly reduced that postcollapse societies are nearly impossible for archaeologists to detect.

7. SYNTHESIS AND IMPLICATIONS

Collapse is a recurrent historical process. It affects the entire range of societies, from those that are least complex to the most complex. Perhaps no society is immune to the possibility of collapse, including the very complex societies of today. The misfortune of

some previous societies is potentially our good fortune. By understanding the processes that cause collapse we can develop tools to assess our own vulnerability.

Earlier research often pinpointed diminished energy flow as the cause of collapse. In many cases, though, diminished energy flow was the product of human management. This is evident in the examples of the Third Dynasty of Ur and the Abbasid Caliphate. In both cases, excessive irrigation caused salinization that destroyed productive soils. In these cases energy does not suffice to explain collapse. Other factors were involved, particularly the tendency of rulers to demand more than the land could sustainably produce. When we ask why the rulers did this, we realize that energy is but a small part of explaining these collapses. The Maya drastically altered their environment, shortening food chains and diverting large parts of the region's productivity to support dense human populations and their complex societies. Yet demands to support complexity so stressed the population that the system collapsed, the population plummeted, and the region reverted to rainforest. Again we see the interplay of energy with politics and human strategies of survival and competition.

Problem solving links these examples to a more general understanding of collapse. Moreover, problem solving is the key to using past collapses to understand our situation today. As noted previously, sustainability is a function of problem solving, as is its opposite, collapse. As the problems that a society or institution faces grow in size and complexity, problem solving grows in complexity and costliness. Problem solving is an economic process, and its success depends on the ratio of the benefits of problem solving to its costs. Problem solving tends to develop from solutions that are simple and inexpensive to those that are increasingly complex and costly. Each increase in complexity has costs, which may be measured in labor, time, money, or energy. Growth in complexity, like any investment, ultimately reaches diminishing returns, at which point increments to complexity no longer yield proportionate returns. When this point is reached, a society or other institution begins to experience economic weakness and disaffection of the populace. A fundamental element of a sustainable society is that it must have sustainable institutions of problem solving. These will be institutions that give stable or increasing returns (e.g., by minimizing costs) or diminishing returns that can be subsidized by energy supplies of assured quality, quantity, and cost.

Ancient societies, constrained by photosynthesis, lacked the energy resources to pay for ever-increasing complexity. The problem persists among today's complex societies, which can pay for complex problem solving only as long as energy is abundant and inexpensive.

The role of problem solving is especially clear in the collapse of the Roman Empire. Confronted in the third century AD with severe challenges to its existence, the empire responded by greatly expanding the complexity and size of its primary problem-solving institutions: the army and government. This was costly and had to be paid for by greatly increasing taxation. The costs were so high that the empire survived by consuming its capital resources: producing lands and peasant population. This was an unsustainable strategy and collapse was inevitable. All that the empire could accomplish was to postpone it. Collapse resulted from the economic consequences of diminishing returns to complexity in problem solving. Ironically, though, had the empire not taken the steps it did, it would have collapsed even sooner.

Problem solving is the element missing in earlier collapse studies that focused unsatisfactorily on energy alone. A prolonged period of diminishing returns to complexity in problem solving renders a society fiscally weak (Figs. 4 and 5) and causes disaffection of the support population. Both factors render a society more vulnerable to stresses than would otherwise be the case. Thus, a fluctuation in energy flow that a fiscally prosperous society would survive can cause the collapse of a society weakened by the cost of great complexity in problem solving. This helps to clarify why changes in climate can have different effects in different circumstances. A society weakened by the costs of problem solving may be vulnerable to a change in climate that in other circumstances may be survivable.

Problem solving links the experiences of the past to understanding our own circumstances. The economics of problem solving, from less to more complex and from increasing to diminishing returns, constrain all institutions, including those of today. An institution that experiences a prolonged period of diminishing returns to complexity in problem solving will find itself in fiscal distress and ineffective at solving problems. If such an institution is part of a larger system that provides its funding (such as a government agency funded by appropriations), it may be terminated or require ever-larger subsidies. If an entire society finds itself in this position, as ancient societies sometimes did, the well-being of its

populace will decline and it may be unable to surmount a major challenge. Either weakness can make a society vulnerable to collapse.

How might this dilemma make itself felt today? As an example, consider the relationship of global climate change to complexity in problem solving. Nothing before in history equals contemporary global change in magnitude, speed, and numbers of people affected. Our response will involve changes in the complexity of technology, research, and government. We need much research to comprehend climate, the economic dislocations that it will entail, and the new technologies that we will need. We will experience increasing government centralization and regulation. In research, technology, and regulation, and in resettlement of people living in low-lying areas, global change will require societal expenditures that are currently beyond calculation. As discussed previously, past societies were made vulnerable to climate change when such a perturbation came after a period of diminishing returns to investment in complexity. If this is so, it would be wise for us to understand the historical position of our own problem-solving efforts. This is not to suggest that climate change will cause today's complex societies to collapse. What it may do is combine with other challenges to increase the complexity and cost of our problem-solving efforts. Barring dramatic developments that increase the availability and decrease the cost of energy, such developments would reduce fiscal resiliency and living standards and make it more difficult to resolve our problems.

The factors that lead to collapse or sustainability develop over decades, generations, or centuries. We need historical research to understand our position in a problem-solving trajectory. In medicine, for example, we have long been in a phase of diminishing returns in which ever-greater expenditures yield smaller and smaller increments to life expectancy. Once we understand this, it is clear that the costs of health care cannot be contained as long as we demand health and long lives, which we will certainly continue to do. Conversely, our investments in microelectronics appear to be in a phase of increasing returns. This happy state, of course, cannot last forever. A primary element to understanding the sustainability of a society, an institution, or a sector of

investment is to understand long-term changes in the complexity and costliness of problem solving.

Our citizens and policy makers have such confidence in science that a problem such as global climate change can generate complacency. We assume that a technological solution will emerge. Few of us are aware that problem solving can reach the point of diminishing returns and become ineffective. One of our great advantages over societies of the past is that it is possible for us to learn where we are in the long-term historical patterns that lead to sustainability or collapse. The historical sciences are an essential aspect of being a sustainable society.

SEE ALSO THE FOLLOWING ARTICLES

Cultural Evolution and Energy • Easter Island: Resource Depletion and Collapse • Economic Thought, History of Energy in • Transitions in Energy Use • War and Energy • World History and Energy

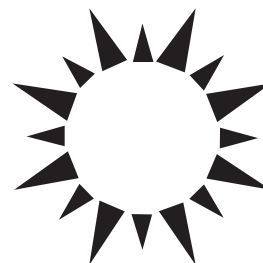
Further Reading

- Adams, R. McC. (1981). "Heartland of Cities." Aldine, Chicago.
- Allen, T. F. H., Tainter, J. A., and Hoekstra, T. W. (2003). "Supply-Side Sustainability." Columbia Univ. Press, New York.
- Bell, B. (1971). The dark ages in ancient history: 1. The first dark age in Egypt. *Am. J. Archaeol.* 75, 1–26.
- Culbert, T. P. (ed.). (1973). "The Classic Maya Collapse." Univ. of New Mexico Press, Albuquerque.
- Jones, A. H. M. (1964). "The Later Roman Empire, 284–602: A Social, Economic and Administrative Survey." Univ. of Oklahoma Press, Norman.
- Renfrew, C. (1979). Systems collapse as social transformation: Catastrophe and anastrophe in early state societies. In "Transformations: Mathematical Approaches to Culture Change" (C. Renfrew, and K. L. Cooke, Eds.), pp. 481–506. Academic Press, New York.
- Tainter, J. A. (1988). "The Collapse of Complex Societies." Cambridge Univ. Press, Cambridge, UK.
- Tainter, J. A. (1999). Post-collapse societies. In "Companion Encyclopedia of Archaeology" (G. Barker, Ed.), pp. 988–1039. Routledge, London.
- Tainter, J. A. (2000). Problem solving: Complexity, history, sustainability. *Population Environ.* 22, 3–41.
- Yoffee, N., and Cowgill, G. L. (eds.). (1988). "The Collapse of Ancient States and Civilizations." Univ. of Arizona Press, Tucson.

Solar Cells

RICHARD CORKISH

University of New South Wales
Sydney, New South Wales, Australia



-
1. Introduction
 2. How Solar Cells Work
 3. Efficiency
 4. Types of Solar Cells
 5. Conclusions

Glossary

diffuse insolation Solar radiation received from the sky, excluding the solar disk.

direct (beam) insolation Solar radiation received directly from the visible solar disk.

efficiency Fraction of incident light power converted to electrical power by a solar cell or module.

global insolation The sum of direct and diffuse insolation components.

insolation Flux of radiant energy from the sun, measured as power per unit area, whose intensity and spectral content varies at the earth's surface due to time of day, season, cloud cover, moisture content of the air, and other factors, but is constant outside the atmosphere at 1353 W m^{-2} .

peak watt (Wp) Rating for a cell, module, or system that produces 1 W electrical output under standard peak insolation conditions and at a standard cell or module temperature, usually 1000 W m^{-2} with global AM1.5 spectrum at 25°C for terrestrial applications, intended to simulate midday conditions at middle latitudes.

photovoltaic effect Generation of an electromotive force (voltage) due to incident photon radiation.

solar module Protectively packaged set of solar cells, electrically interconnected (usually several cells connected in series) to provide useful voltage and current output.

solar spectra Sunlight power content as a function of light color (of radiation frequency) and for which there exist several "standard" representations to aid calculations and comparisons, particularly the air-mass 1.5 (AM1.5) spectrum, which models transmittance through 1.5 atmospheric thicknesses for terrestrial modeling, air-mass zero (AM0), which models extraterrestrial sunlight, and ideal black-body spectra that model the sun as

a thermodynamic black body at temperatures of 6000 K or less.

Solar cells are semiconductor devices that exploit the photovoltaic effect to directly create electric current and voltage from the collection of photons (quanta of light). They convert sunlight to electricity silently and without moving parts, require little maintenance, and are reliable. They are sold with warranties of up to 30 years, generate no greenhouse gases in operation, and are modular, rapidly deployable, and particularly suited to urban rooftops and façades. Solar cells power anything from lanterns in remote villages; to watches, calculators, solar home systems, and grid-connected rooftop arrays; to satellites and exploration vehicles on Mars.

1. INTRODUCTION

According to the review of the field edited by Archer and Hill, solar cells of 15% efficiency covering an area equivalent to just 0.25% of the global area under crops and permanent pasture could meet all the world's primary energy requirements, when most or all of that area would be otherwise alienated land. However, solar cells remain an expensive option for most power-generation requirements relative to fossil and nuclear sources, especially when the natural environment is attributed little or no value, and to some other sustainable options such as wind energy or energy efficiency enhancement.

There is a wide range of solar cells materials and approaches in production, under development, or having been proposed. One way to understand the range is to see the field as comprising three generations of technology. The first includes "bulk" or "wafer" cells, made from layers of semiconductor material thick enough to be self-supporting, usually up to 0.5 mm. This generation, which dominates

markets, is exemplified by single and multicrystalline silicon cells and commercial production cells, which currently have efficiencies of 10 to 15%. Cost reductions for such cells will eventually be limited by the costs of the wafers. The second generation aims for lower cost at the expense of efficiency by using thin layers of active material. A rigid substrate or superstrate provides mechanical support for the semiconductor, which is deposited rapidly over large areas. It is hoped that eventually the relatively low material cost of the substrate or superstrate will dominate the cost of the complete device. Unlike the wafer-based cells described earlier, which are mechanically and electrically connected together to form weatherproof solar modules that produce convenient voltages and currents, the thin film materials are deposited on large areas of sub/superstrate and subsequently patterned to form multiple cells. Hence, the basic production unit is the module, rather than the individual cell. The four important thin-film technologies are amorphous silicon, polycrystalline silicon, cadmium telluride, and copper indium diselenide. Current production efficiencies are 4 to 9%. Thin-film cells have been expected to overtake the first-generation cells' dominance for decades, but this has not yet occurred. In fact, the evolution may be faltering since one major cell manufacturer ceased production of its thin-film products late in 2002 and others struggle for financial viability. Third-generation approaches intend to eventually combine high efficiency and low cost. One type of third-generation cells, tandem cells, are already commercially available as both low-cost, low-efficiency amorphous silicon-germanium and, at the other performance extreme, ultra-high efficiency cells made from elements of Groups III and V of the Periodic Table (known simply as III-V cells).

Market surveys estimated world production to be around 560 MW_p in 2002, an estimated 39% increase over the previous year. Ten companies accounted for 84% of global production in 2002. The three largest were Sharp, BP Solar, and Kyocera, in order of production volume. Sharp's rapid expansion during 2002 gave it nearly twice the capacity of its next largest rival. Figures 1 and 2 indicate the history of regional and global production levels and the end uses.

One figure of merit on which solar cells and modules are compared is their efficiency. Record efficiencies for research cells and modules using different technologies are published regularly by the journal *Progress in Photovoltaics*. Another more commercially important but less precisely defined

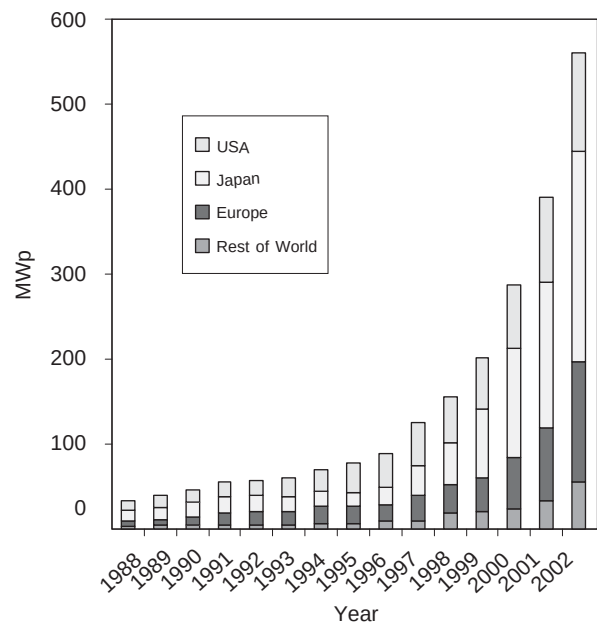


FIGURE 1 World solar cell production, 1988–2002, for the three main production zones and for the rest of the world (ROW). Data from *Photon International* and *Renewable Energy World*.

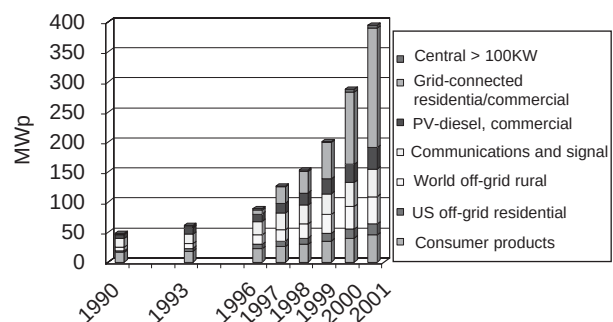


FIGURE 2 World PV market by application, 1990–2001. Data from *Renewable Energy World*.

figure of merit is the cost per W_p. This recognizes that efficiency is only of prime importance in some applications, and for many potential users, the capital cost and the cost of the resultant electricity are more useful measures on which to base decisions. Hence, technologies that are less efficient but cheaper per unit area are able to compete in the market.

2. HOW SOLAR CELLS WORK

Solar cells use semiconducting materials to achieve two basic tasks: (1) the absorption of light energy to generate free charge carriers within the material and (2) the separation of the negative and positive charge

carriers to produce unidirectional electrical current through terminals that have a voltage difference between them. The separation function is usually achieved by a p-n junction, the interface of regions of material that have been “doped” with different impurities to give an excess of free electrons (n-type) on one side of the junction and a dearth of them (p-type) on the other. The interface region therefore has a built-in electrostatic field that sweeps electrons one way and holes the other. In other words, a solar cell is an illuminated, large-area diode.

Figure 3 indicates the main processes. Photons enter the cell volume through the front surface. High-energy (blue or violet) photons are absorbed strongly close to the cell surface, but those from the other end of the visible spectrum are weakly absorbed and penetrate more deeply. Each absorption of a photon annihilates the photon and transfers its energy to an electron in the semiconductor, freeing the electron from its parent atom and leaving behind a positively charged vacancy, or “hole.” In some materials, known as direct bandgap semiconductors, the absorption involves just the photon and the created carrier pair, but in others, indirect bandgap semiconductors, the process also needs one or more phonons (quanta of crystal atomic vibration) to take part. Absorption is weaker in the latter semiconductors. Gallium arsenide and silicon are examples of direct and indirect bandgap semiconductors. The electrons and holes are mobile within the cell and will move in response to an electric field (“drift”) or by diffusion to regions of lower concentration. They are also prone to

recombination with each other. Cell theory usually concentrates on minority carriers (electrons in p-type material and holes in n-type) since their flows (currents) principally determine the performance. Electron-hole pairs generated near the junction are split apart by the strong electric field there. Minority carriers are swept across the junction to become majority carriers and each crossing of the junction is a contribution to the cell’s output current. Minority carriers that are generated too far from the junction to be immediately affected by the junction field can be transported to the junction by diffusion due to the minority carrier population being reduced near the junction by the field action there. Metal contacts at the front and rear of the cell allow connection of the generated current to a load. The front contact is normally in the form of a fine metallic grid to reduce blockage of the light’s access to the semiconductor.

However, survival of a minority carrier is tenuous since it is always prone to recombination with one of the surrounding majority carriers. Recombination can be radiative, the inverse of the optical generation process that produced the carrier pair, in which the carrier energy is lost in the production of a new photon, or nonradiative, in which the energy is dissipated as heat in the cell. Crystal defects, accidental impurities, and surfaces are all strong nonradiative recombination sites. Most nonradiative recombination processes are, in principle, avoidable, but radiative recombination is fundamental.

The result of these processes is a light-generated flow of current in what is normally thought of as the

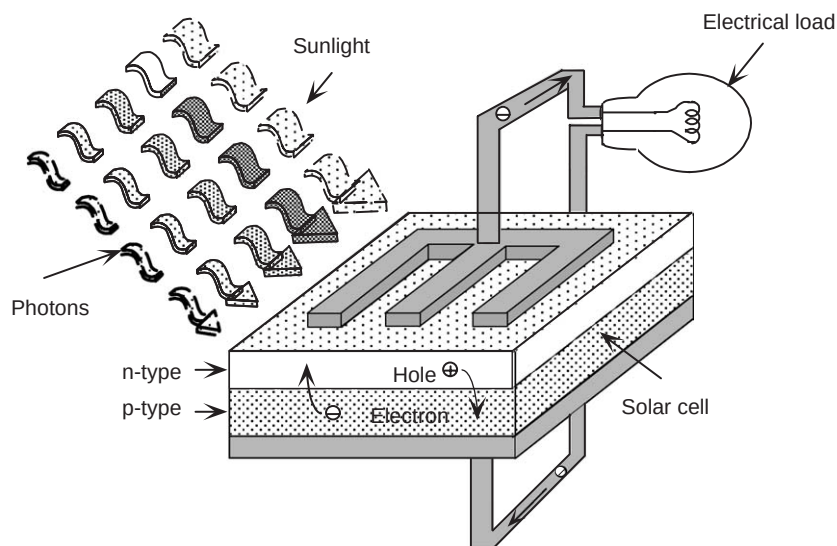


FIGURE 3 Solar cell operation. Adapted from M. A. Green and J. Hansen (2003). *Catalog of Solar Cell Drawings*. University of New South Wales.

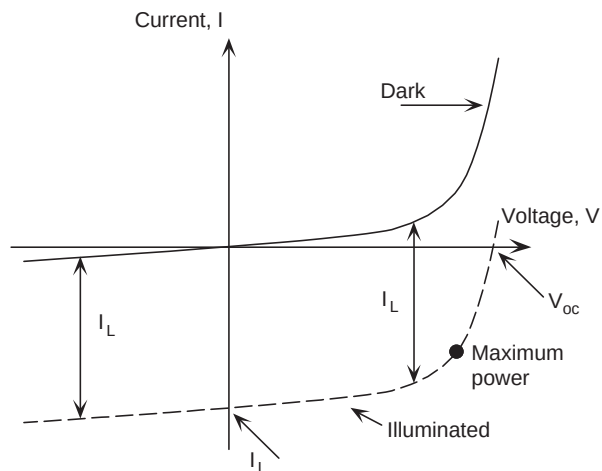


FIGURE 4 Typical current-voltage characteristic. Adapted from M. A. Green and J. Hansen (2003). *Catalog of Solar Cell Drawings*. University of New South Wales.

reverse direction in diode theory (i.e., electrons flow out of the cell into the circuit from the n-type contact and back into the cell through the p-type contact) even though the voltage across it is in the forward bias direction (i.e., positive at the p-type contact). The light-generated current is approximately independent of the voltage across the cell. Unfortunately, the normal diode current, which would remain even in the dark at the same contact voltage, may be thought of as simultaneously flowing in the opposite direction to the light-generated current. That dark current (solid line in Fig. 4) increases strongly with the voltage and at some voltage, known as the open circuit voltage (V_{oc}), completely cancels the light-generated current. As a result, we see the typical current-voltage characteristic shown as a dashed line in Fig. 4. The (forward) dark current is added to the (reverse) light-generated current to make the resultant current negative for a range of positive voltage. In that range, the cell generates power. Power is zero at zero volts and at V_{oc} with maximum power being produced near the so-called knee of the curve. The electrical load connected to the cell should be chosen to keep the operating point close to the optimum knee point during normal operation.

3. EFFICIENCY

In that brief description of cell operation we considered the main processes occurring after photons enter into the semiconductor volume. However, reflection from the surface must first be reduced to allow efficient absorption. A polished

semiconductor surface will reflect a significant fraction of incident photons from the sun, around 30% for silicon. Additional antireflection layers of other materials can be applied to the front surface. In addition, texturing of the front surface can assist with reflection reduction by deflecting rays of light so that they strike the semiconductor surface again. Texturing also redirects photons into oblique paths within the cell, giving longer paths for weakly absorbed (red) photons and more opportunity for absorption (light trapping).

The bandgap energy of the semiconductor from which a cell is made sets the fundamental upper limit of the cell's conversion efficiency. The choice of a low bandgap material allows the threshold energy to be exceeded by a large fraction of the photons in sunlight, allowing a potentially high current. However, a solar cell can extract from each photon only an amount of energy slightly smaller than the bandgap energy, with the rest lost as heat. On the other hand, a semiconductor is transparent to photons with energy less than its bandgap and cannot capture their energy. These, thermalization and transparency, are two of the largest loss mechanisms in conventional cells. The best solar cell materials have bandgaps giving a compromise between these two effects. Solar cell performance is limited by the laws of thermodynamics. Thermodynamic methods and the related principle of detailed balance can be used to calculate the limits. Shockley and Queisser showed in 1961, by use of the balance between incident and escaping photons and extracted electrons, that the efficiency limit for a single-material cell is around 31% for an optimal bandgap of around 1.1 eV. That work assumed that the only unavoidable losses from the cell are the emission of photons produced by radiative recombination.

For such limit calculations, the sun is commonly assumed to behave as a thermodynamic "black body," or perfect radiator at 6000 K, although this is only a rough approximation to the solar spectrum at the earth. The cell is usually assumed to be another perfect radiator at 300 K (27°C). Other losses are assumed to be, in principle, avoidable or controllable and are neglected. Within that framework, a range of different solar cell concepts produces a range of efficiency limits (Table I) from 31% to 86.8% for maximum concentration, which refers to focusing only the solar disk onto the cell, with a concentration of around 46,000 times.

The Carnot limit for solar conversion under the preceding temperatures is 95%, but this could be achieved only if the cell were to produce infinitesimal

TABLE I

Efficiency Limits for Single and Multiple Bandgap Solar Cells for Unconcentrated Sunlight and for Sunlight Concentrated Maximally^a

Number of bandgaps	Unconcentrated light	Maximal concentration
1	31.0%	40.8%
2	42.9%	55.7
3	49.3%	63.9%
∞	68.2%	86.8%

^aOnly the solar disk is seen by the cell and all emissions from the cell are directed back toward the solar disk. The sun is assumed to be a thermodynamic black body radiator with absolute temperature 20 times that of the cell (e.g., cell and sun temperatures of 300 K and 6000 K). From De Vos (1992).

electrical output since the cell would need to emit light at the same spectrum and intensity as the sun in order to avoid entropy production in the energy transfer. The Landsberg efficiency for direct sunlight conversion (93.3%) is the ultimate solar conversion efficiency and results from the assumption that the cell generates no entropy during absorption and emission of light but is lower than the Carnot limit because the light emitted by a cell is considered lost, not returned to the sun for later use. Inclusion of those additional entropy production mechanisms yields, for a multibandgap converter with an infinite number of bandgaps, the 86.8% value in Table I.

4. TYPES OF SOLAR CELLS

4.1 Silicon Cells

Silicon enjoys the advantages of being nontoxic and of great abundance, and its native oxide, silicon dioxide, is easily able to passivate the surfaces to improve the electronic properties there. Silicon solar cell research takes advantage of the vast body of research carried out for the silicon-based electronics industry and, until recently, off-specification silicon rejected by the electronics industry supplied most of the feedstock for the solar cell industry. In addition, useful chemical techniques, such as directionally selective etches, are available for monocrystalline silicon processing. As a result, silicon has dominated the solar cell industry since 1954 and continues to do so. Silicon is used in monocrystalline, multicrystalline, polycrystalline thin-film, and amorphous forms for different types of cells, with the first two supplying most of the market at present.

4.1.1 Silicon Wafer Cells

4.1.1.1 Monocrystalline Silicon Large (up to 300 mm diameter), cylindrical ingots of extremely pure, single-crystal silicon are grown from molten silicon. The entire ingot is doped, usually p-type with boron, during the melt phase and is sawn into circular wafers less than 0.5 mm thick, from which solar cells are made. According to *Photon International*, such monocrystalline silicon cells supplied 36% of the terrestrial PV market in 2002.

Different manufacturers use variations of the same basic process for making screen-printed cells (Fig. 5). Wafer surfaces are etched to form textured surfaces for antireflection and light trapping. Light trapping is implemented to increase the path length within the cell volume for weakly absorbed light at the red end of the spectrum. An n-type dopant, usually phosphorus, is then diffused into the surface to form the p-n junction, producing the in-built electric field to separate the charge carriers (Fig. 3). Heavily doped layers are formed at each surface to aid electrical contact and the metal contacts are formed by the screen printing of a metallic paste before firing. Optical coatings may be added to the front surface to further reduce reflection.

Production cells, using the normal boron doped, solar-grade silicon wafers grown by the Czochralski technique, have efficiencies of typically 12 to 15% although elevated temperatures during operation lead to reduced output power, characterized by a temperature coefficient in the range -0.4 to $-0.5\%/^{\circ}\text{C}$. Much higher efficiency, up to 24.7%, has been achieved in laboratory cells that use expensive photolithographic processes and purer float-zone wafers in their production, but such cells are only commercially viable for high-value applications.

4.1.1.2 Buried and Point Contact Cells An alternative production sequence, used commercially by BP Solar at their Spanish factory, is the buried contact solar cell (Fig. 6). Front contacts are plated into grooves in the cell formed by laser or mechanical grooving, allowing lower metallic shading of the surface while still permitting sufficient metal cross section to have low electrical resistance. BP Solar have reported that processing costs per unit area are only slightly higher than those of the screen-printed cell, but efficiency is significantly higher. These cells may be made from either monocrystalline or multicrystalline silicon.

Sunpower (USA) produce point contact silicon cells with front surfaces completely free from metal

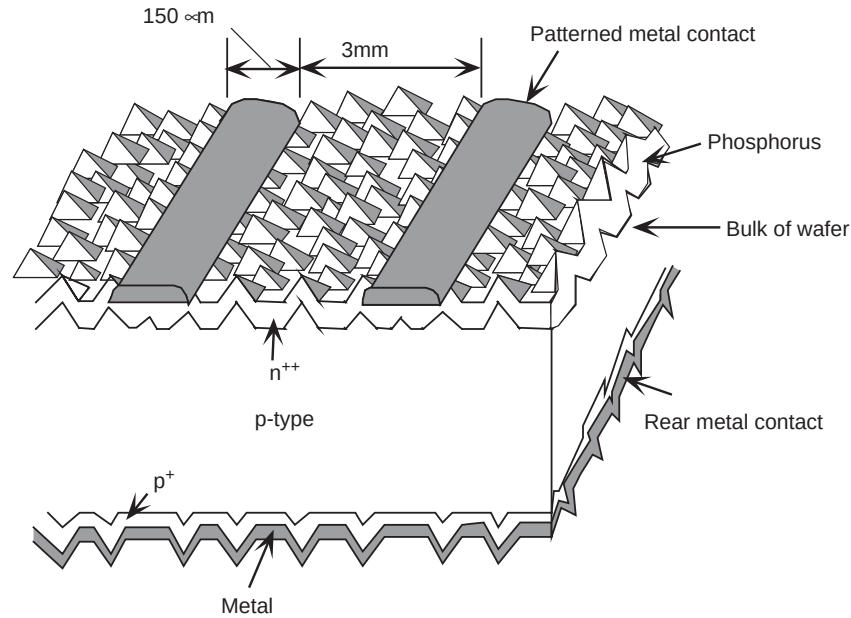


FIGURE 5 Structure of a typical commercial cell with textured surface and screen-printed contacts. From M. A. Green and J. Hansen (2003). *Catalog of Solar Cell Drawings*. University of New South Wales.

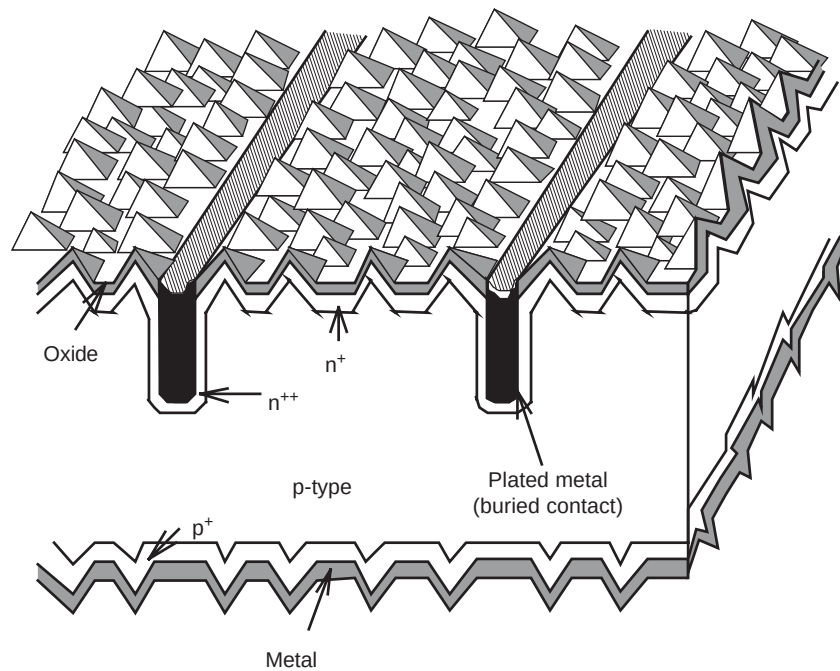


FIGURE 6 Structure of buried-contact silicon cell. From M. A. Green and J. Hansen (2003). *Catalog of Solar Cell Drawings*. University of New South Wales.

contacts. Both the n- and p-type contacts are at the rear. These cells were originally developed for concentrator applications but are predicted to be available in flat plate modules of nominally 20% efficiency from 2004.

4.1.1.3 Semitransparent Monocrystalline Silicon Different approaches have been taken toward making photovoltaic products that are aesthetically attractive for architectural applications. Semitransparency has been achieved with monocrystalline

silicon by, for example, Sunways in Germany. The cells are laser scribed through half the thickness in multiple parallel lines from the front and scribed again orthogonally from the rear. A mesh of holes through the cell is thereby created at the intersections of the lines. The holes and the additional surface area reduce their efficiency, but 9.0% cell and 7.4% module efficiencies are still achieved. Australian National University announced late in 2002 the development and imminent commercialization of a new monocrystalline silicon cell process that uses tiny interconnected cells about 1 mm wide, cut from 1-mm-thick wafers, with light trapping between lower and upper glass sheets.

4.1.1.4 HIT Sanyo's HIT (heterojunction with intrinsic thin-layer) cell, shown schematically in Fig. 7, combines crystalline and amorphous silicon (amorphous silicon is discussed below). n-type crystalline silicon wafers have layers of intrinsic and doped amorphous silicon applied to the front and rear. The junction has intrinsic amorphous layers between the n-type crystalline and p-type amorphous silicon regions to reduce recombination. Contacts are made by metal electrodes and layers of transparent conducting oxide (TCO). Unwelcome absorption in the TCO and electrically inactive doped amorphous layers is compensated by the excellent passivation of the crystalline silicon by the amorphous. Their negative temperature coefficient ($-0.33\%/^{\circ}\text{C}$) is smaller than that of conventional silicon cells. Large-area laboratory cell efficiency of 21.3% has been reported, and Sanyo's nominal cell production

efficiency is up to 19.5%. Sanyo's HIT200 module has nominal cell and module efficiencies of 19.5% and 17%, respectively. Sanyo produced about 16 MW_p during 2001 and reportedly intends to increase production from to 120 MW_p per annum by 2005.

4.1.1.5 Cast Multicrystalline Silicon Cheaper and less energy-intensive wafers are produced by casting (cooling molten silicon in a mould) and sawing. These materials are produced specifically for PV production and have the advantage that the usually square wafers make cells that pack together more closely in modules, leaving less unproductive space between. The resulting material is of lower quality due to the poor electronic qualities of the boundaries between the randomly oriented silicon grains (crystals). Cast and deposited silicon is denoted as multicrystalline, polycrystalline, microcrystalline or nanocrystalline, depending on the average sizes of the crystals. An additional cause of reduced cell performance is the difficulty in texturing such material, and commercial multicrystalline cells are commonly untextured. The use of a passivating/antireflection coating such as silicon nitride is therefore common. Otherwise, processing is similar to that described earlier for monocrystalline silicon cells, and multicrystalline material is used for both screen-printed and buried-contact cells. In 2002, 52% of global production was of cast multicrystalline cells.

4.1.1.6 Ribbon and Sheet Silicon Various methods are used to directly produce silicon sheets of thickness suitable for production of self-supporting

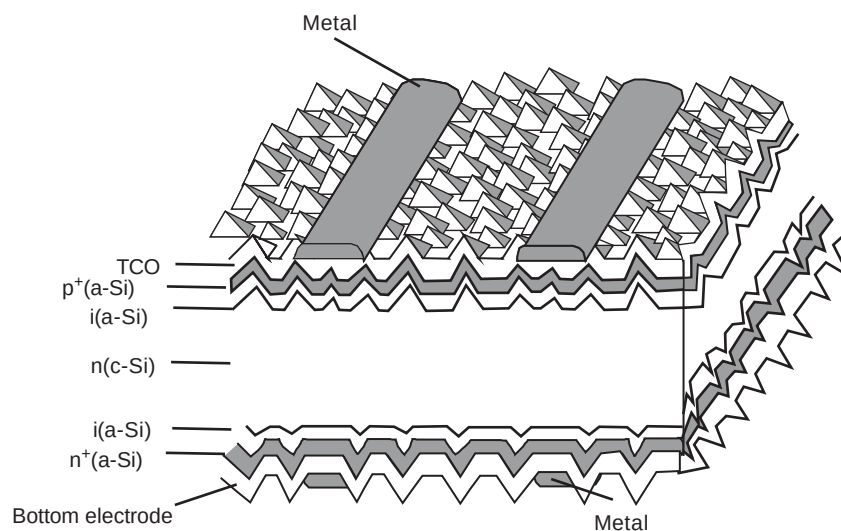


FIGURE 7 Structure of HIT cell. From M. A. Green and J. Hansen (2003). *Catalog of Solar Cell Drawings*. University of New South Wales.

cells without the wasteful sawing process. Edge-defined film growth (EFG) cells and modules are produced by RWE Schott Solar in the United States. A large, hollow nonagon of silicon is drawn through a graphite dye, and wafers are cut from each side. The material is multicrystalline with large, elongated grains and the resulting cells have efficiency exceeding 14% in production and 15 to 16% for best cases. Evergreen Solar grow cell material by a so-called string ribbon process, achieving 15 to 16% best-cell efficiency at standard temperature. Molten silicon forms a film between two carbon strings as they are withdrawn from the crucible. The company was reported in 2003 to be seeking investment to double capacity to 14 MW_p per year. Alternative processes, dendritic web and RGS silicon foil are also proceeding toward pilot production. Ribbon or sheet crystalline silicon cells accounted for 4.7% of world production in 2002, reduced from 5.6% in 2001.

4.1.2 Detached Silicon

Several approaches have been developed to grow thin crystalline silicon films on substrates from which they are detached and the substrates reused. These are aimed at greatly reducing the silicon used, saving a fraction of the necessary investment of energy and money. One, developed by Mitsubishi, involves deposition and laser-crystallization of silicon on an oxidized silicon wafer after holes have been etched through the oxide. More silicon is then grown epitaxially to increase the layer thickness prior to detachment. Another approach is to form a porous silicon layer on a wafer and use that as a crystalline template for growth of a silicon layer before detaching at the porous layer.

4.1.3 Spherical Silicon Cells

Spherical cells are currently being commercialized by Spherical Solar Power (Canada) in partnership with Photowatt (France) and by Kyosemi Corp. (Japan). Small silicon spheres, each a tiny cell of 0.5 to 2 mm diameter, are interconnected in sheets. This results in a flexible and lightweight cell structure, which uses significantly less silicon than conventional wafer cells.

4.1.4 Amorphous Thin-Film Silicon and Silicon-Germanium Cells

In amorphous silicon (a-Si), the silicon atoms are arranged more randomly than in the strict geometric order found in crystals. Many atoms have bonds that are unsatisfied by neighboring silicon atoms so hydrogen is usually incorporated into the material

to reduce recombination, forming a-Si:H. Its absorption coefficient is higher than that of crystalline silicon, allowing it to absorb more strongly and to require lower film thickness ($\sim 0.4 \mu\text{m}$). Relative to wafer-based silicon, thin-film (second-generation) amorphous cells have the potential for lower cost per unit area but also lower efficiency.

However, a light-induced degradation effect has severely restricted the early promise of this technology. Degradation has been minimized by making the cells as thin as possible, and textured TCO surfaces have been used to recover the absorptivity lost by thinning. Germanium is sometimes alloyed with the silicon to modify the bandgap energy since amorphous silicon's bandgap of around 1.7 eV is higher than optimal.

a-Si:H is usually applied by plasma-enhanced chemical vapor deposition. Amorphous cells use a p-i-n structure, with a layer of intrinsic (undoped) material between the n- and p-doped layers to take advantage of the better properties of intrinsic material. In the p-i-n design the junction electric field extends across the intrinsic region, permitting more thickness for generation and collection within the field region. Up to three junctions can be stacked in series for higher output, and triple-junction tandem modules (discussed later) are available commercially. There are two main motivations for this structure: first, the very thin layers required by stability considerations are too thin to absorb all the sunlight; second, it allows the opportunity to have a different bandgap in each junction by alloying a different fraction of germanium in each, giving a better match to the solar spectrum.

Amorphous modules have only a small negative temperature coefficient of power ($-0.1\%/^{\circ}\text{C}$), making them less subject to output reduction due to heat, and they are more tolerant of reduced intensity and diffuse light. The relatively low-temperature deposition technology allows plastics to be used as substrates in flexible modules. They also lend themselves to use in semitransparent building products and car sunroofs. Amorphous silicon's share of production has been steadily declining, falling to 6.4% in 2002. Best reported stabilized efficiency for an amorphous module is 10.4% for a triple-junction alloy device, and commercial triple-junction modules are around 6.3% efficient.

4.1.5 Micro- and Polycrystalline Thin-Film Silicon Cells

Various methods are being investigated to develop thin-film silicon cells deposited on foreign substrates.

If the ratio of hydrogen to silane in the gas from which amorphous silicon is deposited is increased, the resulting material becomes microcrystalline, with columns of crystallites separated by amorphous regions. The optical and electronic properties are similar to those of bulk silicon. Such material has been used as an alternative to silicon/germanium alloys in hybrid structures with amorphous silicon by Kaneka (Japan), who report stable 10% efficiency in a 39W module and who also manufacture amorphous silicon modules. The hybrid modules are suggested to be more stable than amorphous, but that has not yet been independently verified. Particular measures are necessary to allow the amorphous layers to be kept thin enough to avoid light-induced degradation while producing similar current to the microcrystalline cell(s) in series. The University of Neuchâtel, Switzerland, has developed a microcrystalline/amorphous tandem design with around 11% efficiency on a laboratory scale.

Astropower was first to commercially produce thin-film cells with millimetre-sized grains on foreign substrates. Small area cell efficiency of 16.6% has been demonstrated on an unspecified substrate and 9% on a proprietary ceramic. However, the product has evolved to films deposited on a silicon buffer layer, now resembling silicon ribbon cells. Pacific Solar is approaching commercial production with a process in which a thin-film silicon cell is formed on a textured glass superstrate. A laser is used to form craters through the active material to contact the n-type layer closest to the glass. Low-quality material is deposited then improved by subsequent thermal steps. The best efficiency achieved at the time of writing is 8.2% for 30 cm by 40 cm modules, and work is in progress to achieve the same performance from larger modules.

4.2 Thin-Film Cadmium-Telluride Cells

Cadmium telluride (CdTe) thin-film cell efficiency of 16.5% has been achieved, and the record module efficiency currently stands at 10.6%. CdTe comprised 0.7% of global cell production in 2002, and most of that was for indoor use in consumer products. Commercial modules are generally less than 7% efficient. The basic structure has a glass superstrate and a layer of transparent conducting oxide (TCO) as front contact, near-transparent n-type cadmium sulphide (CdS) window layer, p-type CdTe, and metallic rear contact. The contact films are scribed to divide the module into cells, which are series-connected to produce the required voltage.

Potential restrictions on market growth are the toxicity of cadmium and limited global supply of tellurium. Unencapsulated cells are sensitive to moisture. The toxicity issue is controversial and affects public acceptance of the technology. Relatively small-scale production has begun in Japan, the United States, and Germany but production by major players in Germany and the United States ceased in late 2002.

4.3 Thin-Film Chalcopyrite Cells

Copper indium gallium diselenide (Cu(In,Ga)Se_2 , abbreviated to CIGS, or to CIS if the gallium is absent, is another promising thin-film PV material, although it is still only in small-scale production. Laboratory efficiencies have been difficult to transfer to reliable mass production of high-efficiency large modules. Large grains can be produced on foreign substrates, grain boundaries can be passivated, and the material is tolerant of deviations from exact ratios of the constituent elements. Problems include the need to develop specific deposition equipment, the great complexity of co-depositing so many elements, process control, inhomogeneity across large areas, and moisture sensitivity of unencapsulated cells. Global supplies of indium may be a restriction on widespread use in the long term. Toxicity of the cadmium in the window layer may possibly be avoided by use of alternative compounds such as In(OH, S) , although CdS appears to be most commonly used.

A molybdenum rear contact is deposited on a soda-lime glass substrate and patterned by laser to divide into cells. The elements of the CIGS absorber layer are evaporated simultaneously before a heterojunction is formed with a cadmium sulphide layer. A TCO front electrode, mechanical scribing, contact formation, and encapsulation complete the module. One manufacturer expects to achieve better than 8% efficiency in mass production of modules.

Confirmed best cell and module efficiencies are currently 18.4% and 13.4%, respectively. Commercial production in 2002 amounted to just 0.2% of world solar cells. Flexible modules for portable, personal power provision is one high-value market being targeted by CIGS manufacturers.

4.4 Group III-V Cells

Compounds of elements from Groups III and V of the Periodic Table, exemplified by gallium arsenide (GaAs), are used extensively in the optoelectronics

and electronics industries and may be used to make excellent but expensive solar cells

4.4.1 Single Junction III-V Cells

High-efficiency solar cells are produced almost entirely for space applications where low weight, high efficiency, and hardness to cosmic radiation are more important than price. Active semiconductor layers are deposited by a variety of techniques on inactive substrates. High-efficiency single-junction cells have been made from GaAs, GaInAsP, and InP. High-bandgap AlGaAs windows are commonly used to reduce surface recombination.

Gallium and indium are not abundant elements, implying high cost and a limited ability to supply a large fraction of the global energy needs in terrestrial applications, except possibly with high concentration. Best terrestrial cell efficiency is currently 27.6% for a GaAs cell under 255 times concentration.

4.4.2 Tandem and Multijunction III-V Cells

Tandem cells are one approach to exceeding the efficiency limits of single-material cells. They reduce the two main losses mentioned earlier, the thermalization of the excess energy of high-energy photons and transparency to low energy photons. Tandem cells are stacks of p - n junctions, each of which is formed from a semiconductor of different bandgap energy. Each responds to a different section of the solar spectrum, yielding higher overall efficiency. By stacking the component cells in order of decreasing bandgap, as shown in Fig. 8, the light is automatically filtered as each cell extracts photons which exceed its bandgap.

The technique is applied commercially in triple-junction amorphous cells (noted earlier) and in crystalline III-V cells, but using the method with crystalline materials in high-efficiency devices presents several challenges. The materials used in the various junctions must be compatible with each other in several ways over a wide temperature range (defined by the fabrication temperatures), and this severely restricts the options. In particular, the thermal expansion coefficients should be similar and the bandgaps should be a good match to the spectrum. For monolithic devices, the crystal lattice constants must be similar to permit low-defect epitaxial growth.

Tandem cells with up to three bandgaps are commercially available, but they are used almost exclusively for space applications, especially the powering of satellites. There have long been plans for extensive terrestrial use in solar concentrators

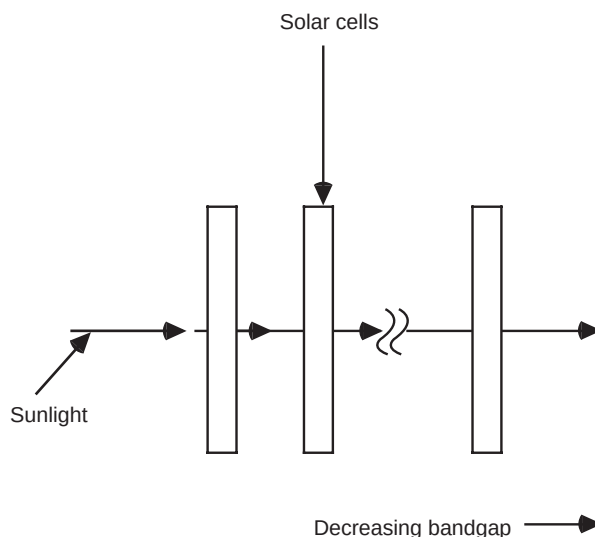


FIGURE 8 Arrangement of bandgaps in a tandem cell. From M. A. Green and J. Hansen (2003). *Catalog of Solar Cell Drawings*. University of New South Wales.

(discussed later), where the concentration of the light from a large collection area onto a cell may be able to justify such expense, but this has not yet occurred to a significant extent. The overall record cell efficiency of 35.2% is for a triple stacked GaInP/GaAs/Ge tandem cell operating under 66-times concentration. Work in Germany during 2003 achieved the first five-junction tandem cell.

4.5 Dye-Sensitized Cells

Dye-sensitized nanocrystalline cells are fundamentally different from the others discussed here in that they do not rely on semiconductor p - n junctions. Instead, they are electrochemical devices in which the optical absorption and carrier-collection processes are separate. A porous film of a wide-bandgap semiconductor, usually titanium dioxide (TiO_2), is coated with a redox charge-transfer dye (Fig. 9). The dye is excited by the absorption of a photon and an electron is injected into excited energy level of the oxide, and the dye is regenerated by a liquid electrolyte. The excited electron in the oxide reaches an electrode and is then able to migrate through an external circuit, do work in the load, and pass to the other electrode, where it regenerates the electrolyte. The front contact is formed from a transparent conducting oxide.

The best-demonstrated efficiency for a laboratory device is 8.2%, and reasonable stability under illumination has been demonstrated. Dye-sensitized cells coming into small-scale commercial production

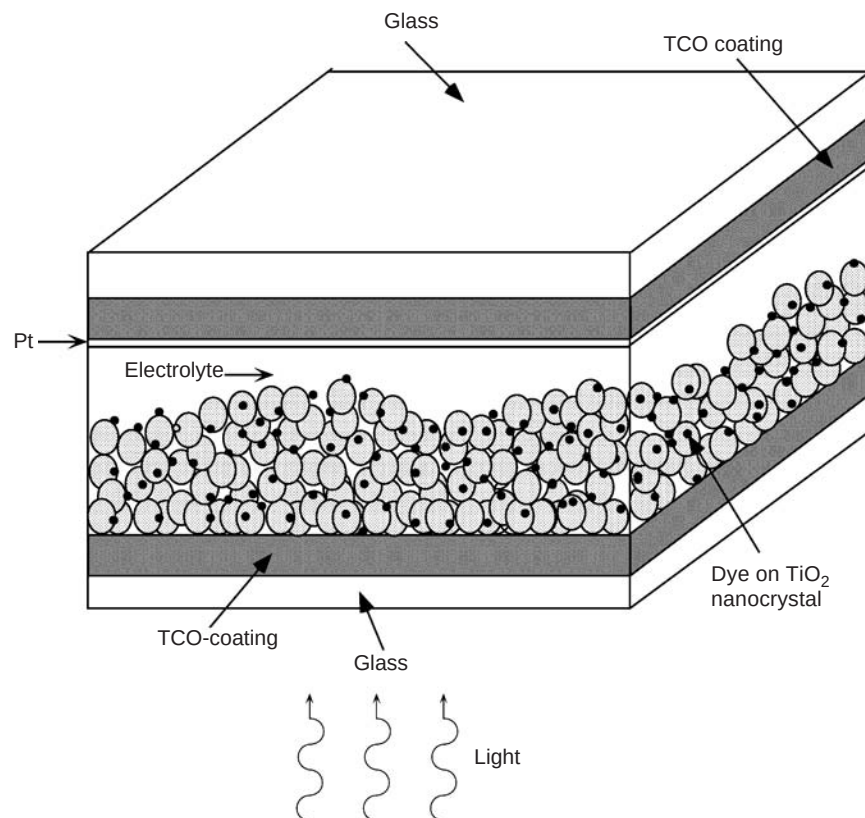


FIGURE 9 Dye-sensitized nanocrystalline cell structure. From M. A. Green and J. Hansen (2003). *Catalog of Solar Cell Drawings*. University of New South Wales.

in Australia, and the installation of the first wall panel has been announced. Dye-sensitized cells exhibit promise as potentially low-cost cells, in terms of finance and energy. Their temperature coefficient can be positive, giving increased power at elevated operating temperatures. Main concerns are to improve high temperature stability and to scale up production. A solid electrolyte would be preferable.

Tandem dye-sensitized cells are aimed particularly at the direct splitting of water for hydrogen production rather than at electricity production. In one experimental system, yielding 4.5% conversion efficiency from sunlight to chemical energy, two semiconductors, TiO_2 and WO_3 or Fe_2O_3 , are used with the water being the electrolyte.

4.6 Concentrator Cells

Sunlight may be optically concentrated to reduce the cell area and reduce costs. Efficiency increases with concentration but at the expense of acceptance of only the direct fraction of the sunlight. Hence, the solar resource is reduced, especially in cloudy or coastal regions. In addition, the concentrating optics

are always less than 100% efficient. Concentrators, other than static concentrators, must track the sun across the sky. They often take the form of Fresnel lenses (point or line focus) or paraboloidal dish or trough reflectors.

Cells can be of either silicon or III-V materials and several commercial products exist, although their undoubted promise has not yet been fulfilled for a variety of reasons. Only a tiny fraction of global production of cells is specifically for concentrators, and one prominent producer of silicon concentrator cells began in 2003 to manufacture flat-plate modules instead. Special care must be taken in the design of concentrator cells, particularly with respect to optics, heating, and high currents. The additional complexity is justified by the reduced cost fraction of the cells in a concentrating system. The heat production in concentrator cells becomes a benefit if it can be economically extracted and there exists a nearby market for low-grade thermal energy. A few combined photovoltaic/thermal systems have been developed.

Few modifications are needed for standard screen-printed silicon cells to operate at low concentration

up to about 10 times. Buried contact silicon cells, for example, operate satisfactorily up to around 30 times concentration, but special cells are required for higher concentration.

4.7 Innovative Solar Cells

In addition to the commercial or near commercial solar cell technologies described earlier, there is a wide array of further ideas for innovative types of cells, aimed at increasing efficiencies or reducing costs or both. Only some of the most prominent, none of which are fully developed and some of which are as yet only ideas, will be mentioned here.

Quantum well cells remain controversial since their proposal in the early 1990s. Multilayer sandwiches of extremely thin alternating layers of high- and low-bandgap semiconductors are held in the static electric field between p and n regions on either side. The layers are thin enough that quantum effects become important, modifying the material properties. The basic intention is that carrier pairs created in the low-bandgap regions will gain sufficient energy optically or thermally to permit their extraction at the higher voltage that is characteristic of the high-bandgap material. Quantum dot cells, in which the quantum effects are evident in three spatial directions instead of one, are a development that promises reduced opportunities for carriers to dissipate their additional energy as heat.

Attempts are being made to apply the impurity photovoltaic and impurity band effects in devices that would be able to utilize two or more photons in multistep absorptions via impurities that introduce energy states within the normally forbidden bandgap. These states also introduce additional recombination pathways, but calculations have shown that in the optimal situation, very high efficiencies would be possible. Making an electrically separate optical upconverter or downconverter by putting the impurities in a separate pieces of material instead of in the cell itself would avoid those recombination paths for some carriers. An all-silicon tandem cell is being pursued through the modification of the silicon bandgap in the upper subcell by forming superlattices of alternating thin silicon layers with layers of silicon dioxide or nitride.

If carriers due to high-energy photons could be collected while still energetic or “hot,” the thermalization loss could be avoided and that is the aim of researchers investigating hot carrier cells. Carriers either have to reach the contacts rapidly or thermalization must be slowed. In addition, development of

contacts that select the energy of the collected carriers is required.

All the approaches mentioned so far in this section are intended to produce very high efficiencies. Organic semiconductor solar cells, on the other hand, are not expected to be as efficient as present commercial cells but are expected to be produced extremely cheaply, at relatively low temperatures over large areas of cheap substrates. These cells take several forms, including interpenetrating networks of two polymers that are chosen to form a heterojunction that separates electrons and holes from each other.

5. CONCLUSIONS

Solar cells technology is evolving through three generations, with the first, comprising cells made from relatively thick (100s of μm thick) wafers of monocrystalline or multicrystalline silicon, continuing to dominate the industry. Silicon has several advantages over competing materials, but ultra-pure semiconductor wafers are inherently expensive in terms of both finance and energy. The second-generation cells—based on thin-films of amorphous, nanocrystalline, or polycrystalline semiconductors, deposited on cheap substrates or superstrates for mechanical support—are entering the market, or, as in the case of a-Si, are already established with a small market share. Second-generation cells may eventually be available at significantly lower cost per unit of delivered energy than first-generation cells, but a range of scientific, materials, production, and environmental issues have hampered the market penetration of the thin-film cells. When their market matures, it is expected that material input costs will eventually dominate their price. However, the production share of thin-film cells is currently declining, while the established first-generation approaches continue to be produced more cheaply due to enhanced manufacturing scale and technical improvements. Silicon may come to dominate the second generation of solar cell production too. Research is already under way into a wide range of attempts to combine the attributes of very high efficiency and low-cost production.

The global market, and particularly in Japan and Europe, is growing rapidly, fuelled by energy security and environmental concerns. Solar cells are complementing wind and waste-derived biomass energy in satisfying the growing demand for energy with low environmental impact. After several false starts, solar cells are finally coming of age.

Acknowledgments

The Centre of Excellence for Advanced Silicon Photovoltaics and Photonics is supported by the Australian Research Council's Centres of Excellence Scheme. The author thanks M. Green and J. Hansen for permission to use Figs. 2 through 9 and to M. Green for helpful comments.

SEE ALSO THE FOLLOWING ARTICLES

Materials for Solar Energy • Photovoltaic Conversion: Space Applications • Photovoltaic Energy: Stand-Alone and Grid-Connected Systems • Photovoltaics, Environmental Impact of • Solar Cookers • Solar Cooling, Dehumidification, and Air-Conditioning • Solar Detoxification and Disinfection • Solar Distillation and Drying • Solar Energy, History of • Solar Fuels and Materials • Solar Heat Pumps • Solar Ponds • Solar Thermal Energy, Industrial Heat Applications • Solar Thermal Power Generation

Further Reading

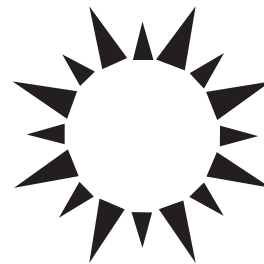
Archer, M.D., and Hill, R. (eds.). (2001). *Clean Electricity from Photovoltaics*. Imperial College Press, London.

- Brendel, R. (2003). "Thin-Film Crystalline Silicon Solar Cells: Physics and Technology." Wiley-VCH, Berlin.
- De Vos, A. (1992). "Endoreversible Thermodynamics of Solar Energy Conversion." Oxford Science, Oxford.
- Grätzel, M. (2001). Photoelectrochemical cells. *Nature* **414**, 338–344.
- Green, M. A. (1986). "Solar Cells: Operating Principles, Technology and System Applications." University of New South Wales, Sydney.
- Green, M. A. (1995). "Silicon Solar Cells: Advanced Principles and Practice, Centre for Photovoltaic Devices and Systems." University of New South Wales, Sydney.
- Green, M. A. (2003). "Third Generation Photovoltaics: Ultra-High Efficiency at Low Cost." Springer-Verlag, New York.
- Green, M. A. (in press). "Thin-Film Photovoltaics: Advances in Solar Energy" (K. W. Böer, Ed.). Amer. Solar Energy Soc., Boulder, CO.
- Luque, A., and Hegedus, S. (eds.). (2003). "Handbook of Photovoltaic Science and Engineering." Wiley, London.
- Markvart, T., and Castañer, L. (eds.). (2003). "Practical Handbook of Photovoltaics". Elsevier, Oxford.
- Shah, A. V., Meier, J., Vallat-Sauvain, E., Wyrsh, N., Kroll, U., Droz, C., and Graf, U. (2003). Material and solar cell research in microcrystalline silicon. *Solar Energy Mat. Solar Cells* **78**, 469–491.
- Tawada, Y., Yamagishi, H., and Yamamoto, Y. (2003). Mass production of thin film silicon PV modules. *Solar Energy Mat. Solar Cells* **78**, 647–662.
- Zweibel, K. (1990). "Harnessing Solar Power: The Photovoltaics Challenge." Plenum Press, New York.

Solar Cookers

S. K. SHARMA

Energy Research Centre
Chandigarh, India



-
1. Introduction
 2. Classification of Solar Cookers
 3. Parabolic Concentrator Cookers
 4. Fresnel Reflector Cookers
 5. Plane Mirror Concentrating Cookers
 6. Transmitting Concentrator Cookers
 7. Fresnel Lens Cookers
 8. Principles of Paraboloid Cookers
 9. Solar Box Cookers
 10. Multiple Glass Box Cookers
 11. Indirect Solar Cookers
 12. Scheffler Community Cooker
 13. Rating of Solar Cookers

Glossary

chemical energy Energy stored in a substance and released when the substance changes form or combines with another substance.

compound parabolic concentrator (CPC) A device that consists of two curved reflecting segments, which are part of parabolas.

concave Surface curved like the interior of a circle or a sphere.

energy Capacity, or ability, to do work.

flux concentration Density of radiation falling on and/or received by a surface or body.

parabola Plane curve formed by the intersection of a cone with a plane that runs parallel to its side.

paraboloid A solid, some of whose plane sections are parabolas.

pyranometer An instrument that measures solar radiation.

radiant energy Energy received in the form of photons or electromagnetic radiation.

rim angle An angle measured at the focus from the axis to the rim or a point where a paraboloid has been truncated.

solar concentrator A device that focuses the solar radiation received on the receiver.

solarimeter An instrument that measures solar radiation.

solar insolation Intensity of solar radiation.

solar radiation A form of electromagnetic radiation.

specific heat capacity The amount of heat energy absorbed or lost by a substance.

thermal energy Also known as heat; the kinetic energy associated with rapid random motion.

1. INTRODUCTION

Cooking is a very important and essential chore in every household. Unfortunately, this activity is also a major cause of deterioration of forest cover and soil fertility. It is also responsible for drudgery and ill health among women in developing countries. The main reason is the use of traditional biomass fuels, such as wood, cow dung, and agricultural residues, in inefficient cookstoves, in poorly ventilated kitchens. It has been estimated that there are 180 polar, 75 aliphatic, and 225 aromatic hydrocarbon compounds in smoke produced from biomass. The exposure of women and children to these carcinogenic, ciliotoxic, and mucus-coagulating agents results in numerous diseases related to the respiratory system, such as colds in adults and acute respiratory infections in infants and young children.

It has been estimated that the use of solar cookers by 2000 families for 10 years would mean a savings of \$0.1 million in terms of reduced firewood consumption, \$2 million worth of coal, or \$0.25 million worth of kerosene oil.

1.1 History

Cooking with energy from the sun is an old concept. There are many landmarks in the history of the design of solar cookers.

The first attempt to cook fruits was made by Swiss scientist Horace de Saussure in 1767. He achieved 88°C in a hot box-type cooker. In 1837, astronomer Sir John Frederick Herschel achieved 116°C in his solar cooker and cooked meat and vegetables. Muguste Mouchot designed a community cooker for baking bread and potatoes for French troops in 1870. The first manual on solar cookers, entitled "Solar Heat: A Substitute for Fuel in Tropical Countries," was written by William Adam in 1878 in India.

Maria Telkes and George O. G. Lof performed extensive studies on solar cooking at the Massachusetts Institute of Technology.

A significant attempt to introduce nearly 200 of the concentrator-type solar cookers developed at the University of Wisconsin was made in the states of Sonora, Coahuila, and Oaxaca, Mexico, with the help of the Rockefeller Foundation. After initial problems, the cooker was redesigned and the modified version of the cookers was used for many years.

One of the largest attempts to introduce solar cookers was made in India in 1953, when 200,000 paraboloid concentrator-type solar cookers developed by M. L. Gai of NPL New Delhi were brought in. The cooker consisted of a paraboloid reflector of equation $Y^2 = 180X$, with a focal length of 45 cm; the reflector was spun from an aluminum sheet into the desired shape and then anodized to protect it from the weather and to maintain reflectivity. The first commercial version of the hot box-type cooker was designed by M. K. Ghosh in India in 1945.

1.2 Cooking Methods and Solar Cookers

The principal methods of cooking food can be divided into four categories based on cooking temperature and cooking media. These are boiling, frying, baking, and roasting.

1.2.1 Boiling

Water is used as a medium in this type of cooking. This sets the upper temperature limit. In this mode of cooking, food is cooked at only 100°C. In cooking by boiling, the mixture of food and water is brought to a boil and allowed to simmer until completion of the chemical reaction of the food. Due to the presence of a large amount of water in this type of cooking, the specific heat of the water and food mixture is assumed to be equal to that of water, i.e., 4 kJ/kg °C. Thus, once the boiling point is reached, not much heat is required except that the heating rate should be equal to the rate of heat loss. The thermal losses from

the cooking vessel area include the following: loss due to vaporization of water (35%), heating of the food–water mixture to the boiling point (20%), and convection losses from the vessel (45%). Box-type solar cookers are suitable for this type of cooking.

1.2.2 Frying

Oil is used as a medium in this type of cooking. The upper temperature limit of this mode is dependent on the characteristics of the oil used. Cooking is dependent on the boiling temperature of oil, which is generally between 200 and 300°C. Frying is a high-power cooking process and cooking is normally completed in a short time. Otherwise, food may become burned. Concentrator cookers are suitable for this type of cooking.

1.2.3 Roasting/Grilling

In this mode of cooking, heat is transferred to the food primarily through radiation and to some extent through convection. It is also a very high power process. Concentrating cookers are suitable for this type of cooking.

1.2.4 Baking

Baking is also a high-power cooking process in which heat is transferred from the oven wall by convection and radiation. Solar ovens are suitable for this type of cooking.

2. CLASSIFICATION OF SOLAR COOKERS

Solar cookers can be classified into four main categories: (1) concentrator cookers; (2) solar ovens; (3) box cookers; and (4) indirect (combined) solar cookers. These main categories can be further subdivided into different categories, as shown in Fig. 1.

3. PARABOLIC CONCENTRATOR COOKERS

A parabolic solar cooker is a concave bowl-shaped dish, whose inner surface is made of reflective material. Sunlight falling on the inner surface is focused onto a dark cooking pot that is suspended or set on a stand in front of the cooker.

The power output of the cooker depends on the size of the dish, the intensity of the solar radiation, the reflectivity of the inner lining, and the perfection

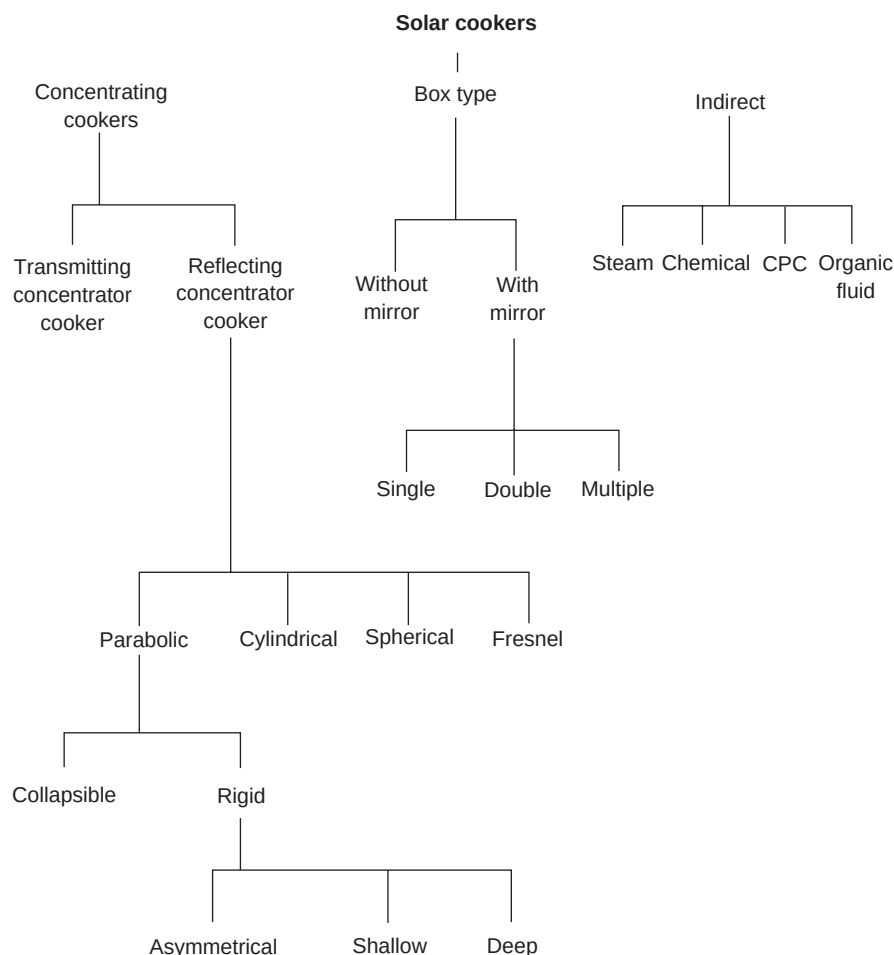


FIGURE 1 Types of solar cookers.

of the geometry. A concentrator cooker having 400 to 600 W output requires an area of 1 m². Concentrator cookers work only in direct sunlight; they work poorly under cloudy conditions. They require constant tracking. They also become unstable under windy conditions.

Maximum numbers of models of this category of concentrating cooker have been developed, as focusing of solar radiation is achieved much better using this geometry. However, constant tracking is required for this type of cooker as these cookers are sensitive to small changes in the relative position of the sun. Fabrication of parabolic geometry is also difficult compared to spherical geometry. In view of this, large sizes are fabricated with small strips and careful calculations are required to arrive at a suitable geometry. Some simpler methods, such as wire-reinforced concrete shells, soil cement depressions in the ground, and papier-maché shells, have been suggested. Parabolic concentrators can be classified

into two categories: rigid parabolas and folding parabolas.

The majority of models that have been developed are based on the rigid parabola design. This category can be further subdivided into the following three subcategories: shallow parabolic cookers, asymmetrical parabolic cookers, and deep parabolic cookers.

3.1 Shallow Parabolic Cookers

In this type of cooker, the focus is outside the rim of the reflector.

A separate cradle-type stand is used to hold the cooking vessel as well as the reflector. In some designs, only the reflector is moved to track the sun.

3.2 Asymmetrical Parabolic Cookers

Asymmetrical parabolas are partial parabolas. This configuration enables the cooking to be done as close

as possible to the vessel. In the former design, there is a hole in the center of the reflector, where a stand is inserted to support the cooking vessel.

In the deep asymmetrical parabolic configuration, several small parabolic reflectors are arranged in an asymmetrical parabolic configuration. See Figs. 2, 3, and 4.

3.3 Deep Parabolic Cookers

In these cookers, the focus is within the rim of the reflector. The cooking vessel is hung from a stand inside the deep parabolic reflectors. Focusing is essential once

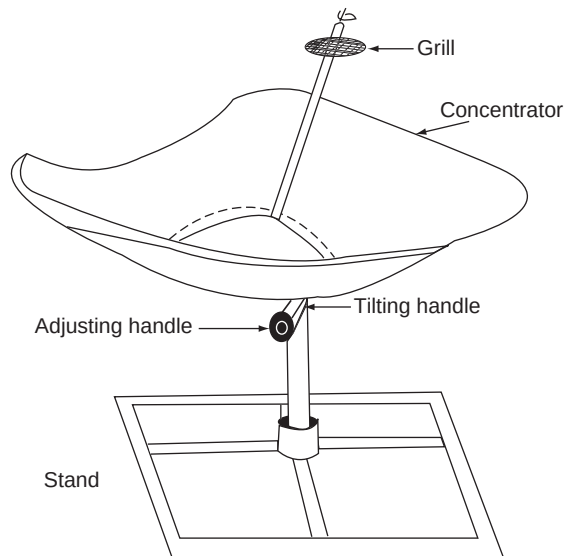


FIGURE 2 Asymmetrical parabolic cooker.

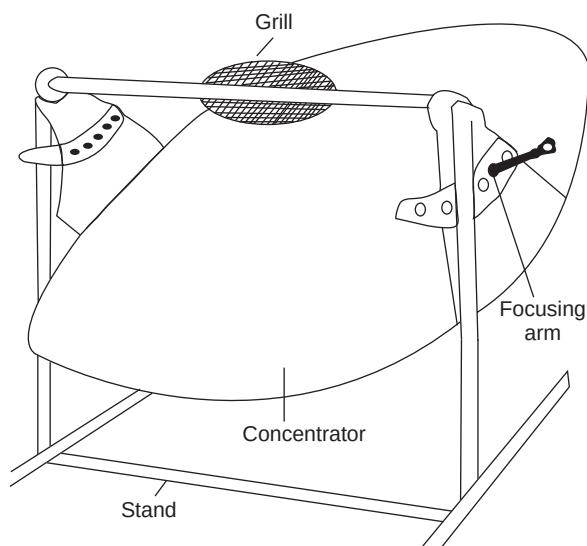


FIGURE 3 Asymmetrical parabolic cooker.

every 20–30 min. These cookers were not very successful, due to bright sunlight hitting the eyes during cooking, the need to handle hot cooking vessels, and the need for periodic tracking. See Figs. 5 and 6.

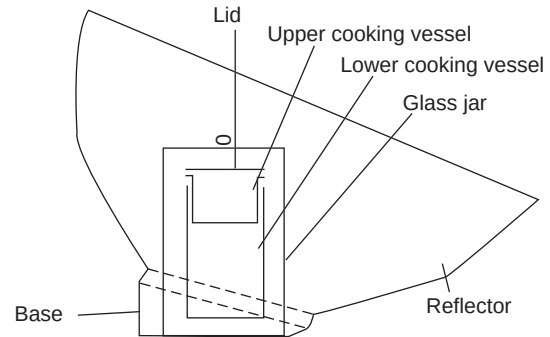


FIGURE 4 Asymmetrical parabolic cooker.



FIGURE 5 Deep parabolic cooker.

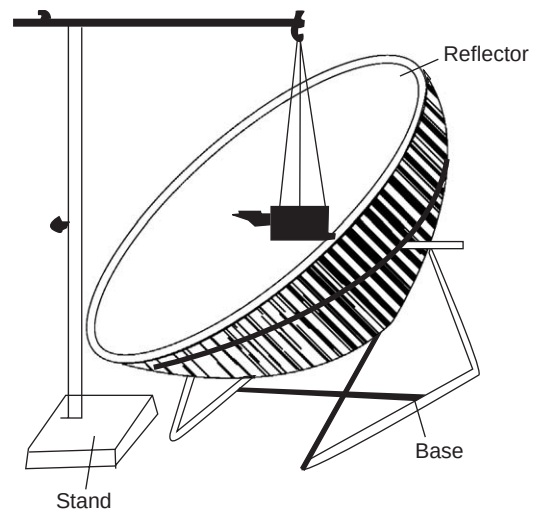


FIGURE 6 Deep parabolic cooker.

3.4 Collapsible Parabolic Cookers

Parabolic cookers are difficult to transport due to their bulky size. To facilitate easy transportation, collapsible parabolic cookers were designed. Some of the designs open like an umbrella. In others, inflatable plastic configurations have been used. Fluttering focus and unstable reflectors were significant impediments in the large-scale propagation of these cookers. See Figs. 7, 8, and 9.

3.5 Cylindro-Parabolic Cookers

In cylindro-parabolic cookers, rays are focused in an insulated box in which the cooking pots are placed. One or more cylindro-parabolic mirrors can be used. The tracking arrangement of these mirrors is such that they focus solar radiation into the cavity of the insulated box. See Figs. 10 and 11.

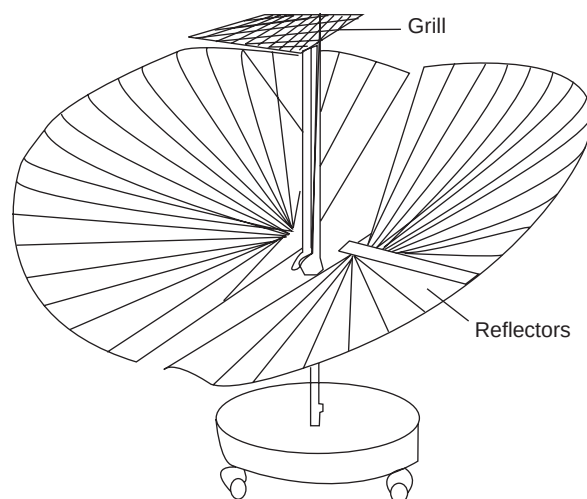


FIGURE 8 Collapsible parabolic cooker.

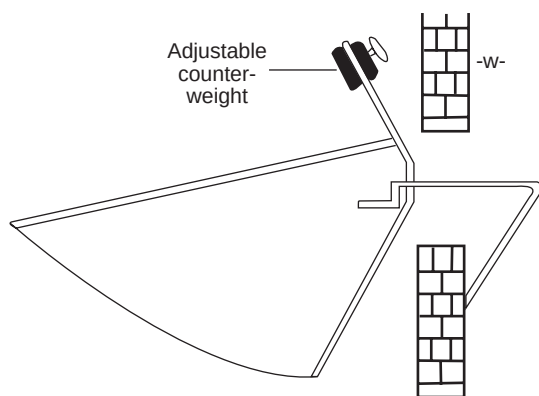


FIGURE 7 Collapsible parabolic cooker.

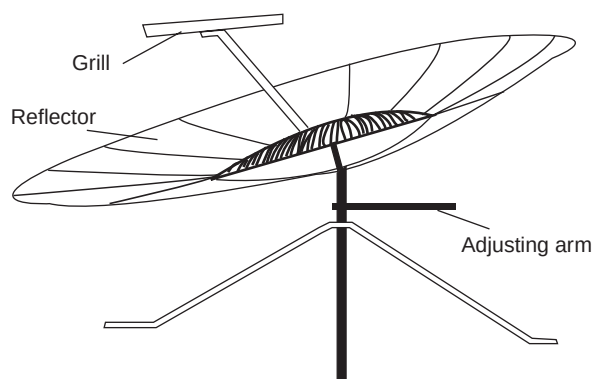


FIGURE 9 Collapsible parabolic cooker.

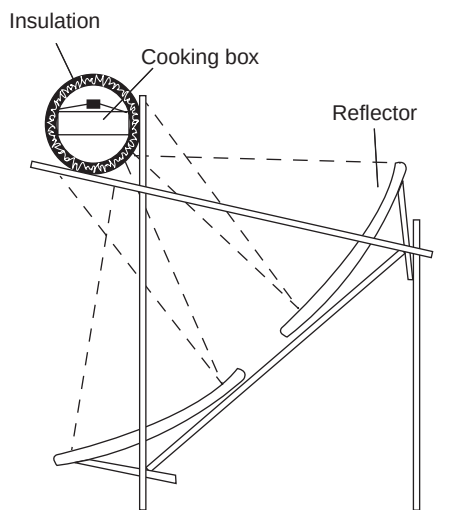
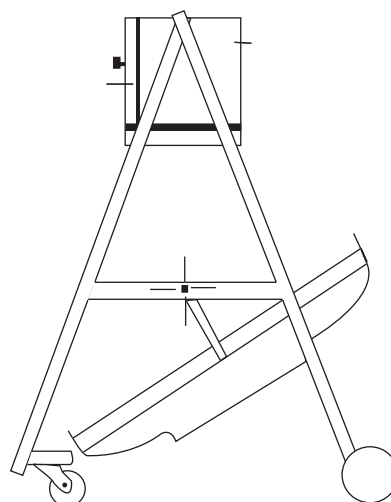


FIGURE 10 Cylindro-parabolic cooker.



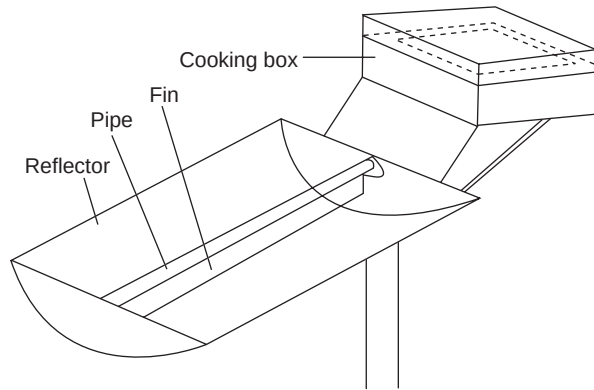


FIGURE 11 Cylindro-parabolic cooker.

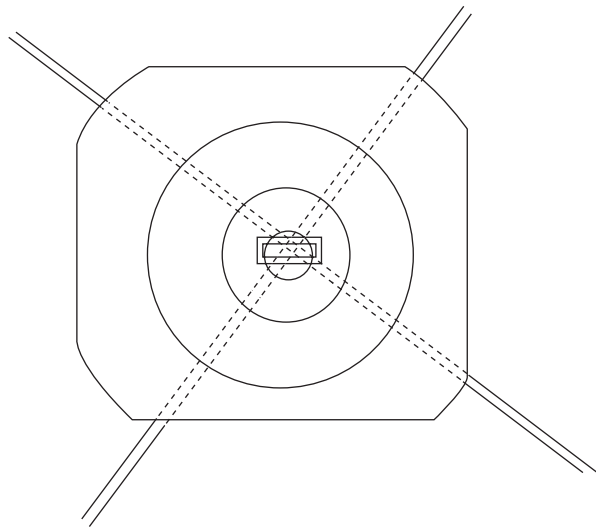


FIGURE 12 Fresnel reflector cooker.

4. FRESNEL REFLECTOR COOKERS

The reflectors for these types of cookers are made by using reflecting rings in a concentric Fresnel geometry. Any size of solar cooker can be made using this technique. However, designing such cookers requires considerable skill. See Fig. 12.

5. PLANE MIRROR CONCENTRATING COOKERS

Plane mirrors have been used in different configurations, such as cones or folding plane stands, such that these mirrors focus rays onto the cooking pot. Provisions are made to tilt the mirror assembly to focus the sun's rays onto the cooking pot during cooking. However, these designs did not become popular due to the large size of the stands. See Fig. 13.

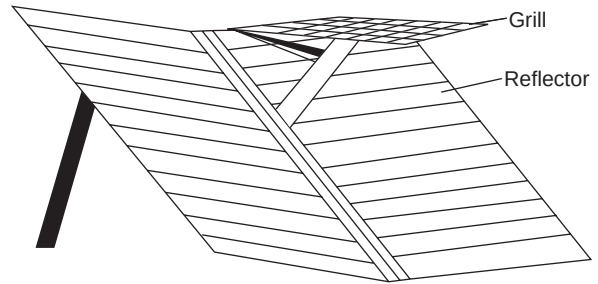


FIGURE 13 Plane mirror concentrating cooker.

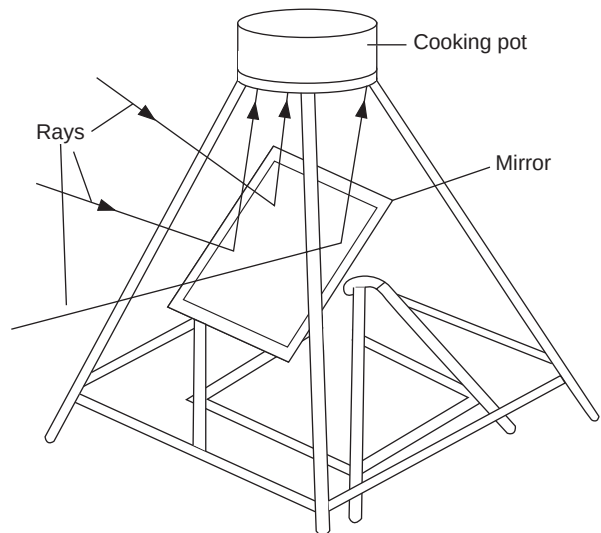


FIGURE 14 Water lens cooker.

6. TRANSMITTING CONCENTRATOR COOKERS

These cookers can be divided into three categories: water lens, dome lens, and Fresnel lens cookers.

6.1 Water Lens Cookers

These types of cookers concentrate light from above. The concentrator in such cookers consists of either plastic or glass concave sheets that are filled with water or alcohol. In some designs, an additional lens of small diameter is used to form a sharp focus. In a modified version, a deflector is used to focus the light onto the bottom of the pot. See Figs. 14 and 15.

6.2 Dome Lens Cookers

In such cookers, a dome comprising multiple cookers is used. However, fabrication of an assembly of lenses having a long focus is difficult.

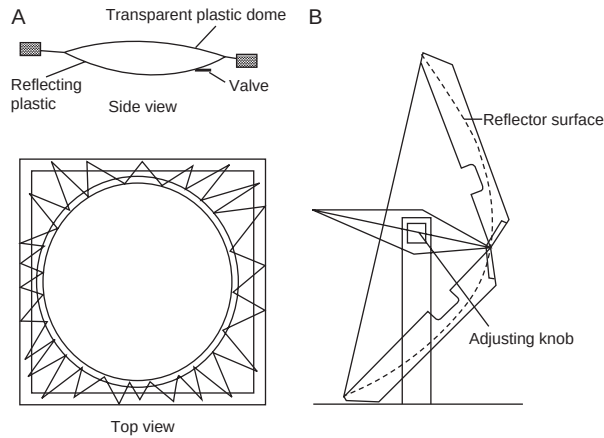


FIGURE 15 Water lens cooker.

7. FRESNEL LENS COOKERS

The use of transparent Fresnel lenses has been suggested for cooking applications. Two configurations of this design have been suggested. In one, the vessel is tracked, and in the other, the lens assembly is tracked. These designs did not become popular due to the high cost of Fresnel lenses. Also, plastic lenses deteriorate very fast. With the availability of ultra-violet-resistant polymers, such cookers may become a feasibility.

8. PRINCIPLES OF PARABOLOID COOKERS

Only the beam component of solar radiation is normally collected by paraboloid concentrating solar cookers, which are reflected to the focal point. Thus, for satisfactory performance, predominantly sunny skies are required. Thermal energy input to the cooking vessel is immediately cut whenever there is an obstruction between the sun and the cooker. The length of time spent in the shade and other factors, such as wind and the size of the cooking vessel, determine the recovery time.

8.1 Theoretical Consideration in the Design of Concentrator Cookers

The basic function of a concentrator is to collect radiation over a large area and concentrate it onto a smaller area so as to attain higher temperatures. Concentration can be defined by the following two methods: geometric concentration and flux concentration.

Geometric concentration is defined as the ratio of aperture area to receiver area

$$C_{\text{Area}} = A_A/A_R. \quad (1)$$

and flux concentration is defined as the ratio of intensity at the aperture to intensity at the receiver

$$C_{\text{Flux}} = I_A/I_R. \quad (2)$$

This involves an absorption effect in addition to geometry.

In concentrator solar cookers, geometric flux is the most widely used method.

According to the second law of thermodynamics, the maximum possible concentration for a given acceptance half-angle, θ_c , for a line-focusing concentrator collector is given by

$$C_{\text{Ideal 2D}} = 1/\sin \theta_c. \quad (3)$$

For point-focusing concentrators (cones, dishes, pyramids), the ideal concentration is given by

$$C_{\text{Ideal 3D}} = 1/\sin^2 \theta_c, \quad (4)$$

where θ_c is the acceptance angle, which is defined as the angular range over which all or almost all rays are accepted without the collector being moved.

8.2 Concentration Limit

The angular radius of the sun $\Delta s = 0.25^\circ$.

Thus, the thermodynamic limit of a tracking two-dimensional solar cooker is 200 and that for a three-dimensional solar cooker is 40,000. However, the actual values achieved are much lower due to the nonoptimized design of the concentrator, tracking errors, nonspecular behavior of the optical system, and atmospheric scattering.

Focal length is an important parameter in the design of concentrator cookers. For spherical concentrator cookers, focal length changes with the length of the curve in a spherical mirror, as described by

$$B = (1/2)r^2/(r^2 - y^2)^{1/2}, \quad (5)$$

where r is radius and y is height.

In parabolic cookers, focal length depends only on the nature of the curve; it does not change with the length of the curve or the point on the curve, as is the case with spherical cookers.

8.3 Performance of Parabolic Concentrators

Parabolic concentrator performance depends mainly on two parameters, namely, the optical efficiency factor, $F\eta_o$, and the heat loss factor, FU_L . The optical efficiency factor depicts the perfection of the reflector surface area, its reflectance, and the absorption of the outer surface of the cooking pot. Thus, it defines the theoretical upper limit of the overall efficiency of a concentrator solar cooker.

The heat loss factor FU_L depends on the temperature of the water in the pot, the wind speed, the surface area of the cooking pot, and the orientation of the reflector. The heat loss factor can be evaluated by studying the cooling curve and optical efficiency can be evaluated by studying the heating curve.

8.4 Heat Loss Factor

8.4.1 Theory

The infinitesimal time interval dt for a fall of infinitesimal temperature of dT_w can be depicted by the following equations:

$$dt = \frac{(MC)\delta_w dT_w}{Q_L} \quad (6)$$

$$= \frac{(MC)\delta_w dT_w}{A_r F\eta_o U_L (T_w - T_a)} \quad (7)$$

Assuming that the heat loss factor U_L and the ambient temperature T_a are constant during the cooling test, the above equations can be integrated over the time interval t during which water temperature falls from $T_{wo} - T_w$.

$$t = \frac{(MC)\delta_w}{A_r F\eta_o U_L} \ln \left[\frac{(T_w - T_a)}{(T_{wo} - T_a)} \right] \quad (8)$$

$$(T_w - T_a) = (T_{wo} - T_a) e^{t/t_o} \quad (9)$$

It can be inferred from the above equations that t_o is time constant, as after time $t = t_o$, the temperature difference $T_w - T_a$ falls to $1/e$ of the initial value.

8.4.2 Optical Efficiency

Optical efficiency can be evaluated by the heating curve method. In this method, the time taken, dt , for a rise in temperature, dT , of water during heating is measured.

$$dt = \frac{(MC)\delta_w dT_w}{Q_u} \quad (10)$$

$$dt = - \frac{(MC)\delta_w dT_w}{F\eta_o [A_p \eta_o I_b - A_t U_L (T_w - T_a)]} \quad (11)$$

If the insulation and ambient temperature are constant over a certain interval of time t , the above equations can be integrated over this time interval when temperature rises from T_{w1} to T_{w2} as

$$t = -t_o \ln \left\{ \frac{f\eta_o - \frac{F\eta_o U_L}{C} \left[\frac{T_{w2} - T_a}{I_b} \right]}{F\eta_o - \frac{F' U_L}{C} \left[\frac{(T_{w1} - T_a)}{I_b} \right]} \right\} \quad (12)$$

where t_o is the time constant obtained from the cooling curve and C is the ratio of aperture area of the cooker to the area of the pot surface (A_p/A_t).

The above equation can be rewritten as

$$F\eta_o = \frac{F\eta_o U_L}{C} \left\{ \frac{[(T_{w2} - T_a)/I_b][(T_{w1} - T_a)/I_b]}{1 - e^{-t/t_o}} \right\} e^{-t/t_o} \quad (13)$$

where $F\eta_o$ is the optical efficiency.

Time to boil can be calculated by writing the equation

$$t_{\text{boil}} = t_o \ln \left\{ \frac{1}{1 - [(F\eta_o U_L)/(F\eta_o)](1/C)[(100 - T_a)/I_b]} \right\} \quad (14)$$

8.4.3 Procedure

The dimensions of the parabolic reflector cooker, such as aperture diameter and rim angle, are measured. The height, diameter, and weight of the cooking vessels are also measured. The cooking vessels are loaded with water and the parabolic concentrator solar cooker is pointed toward the sun in such a manner that the bright spot is positioned on the cooking vessel. A heating curve is obtained until the water temperature reaches 90–95°C. The concentrator is then shaded and a cooling curve is obtained until the temperature is close to ambient.

A semilogarithmic graph between $\ln(T_w - T_a)^\circ\text{C}$ and length of time of cooling is plotted. The slope of the line is equal to $(-1/t_o)$.

The heat loss factor is calculated using the following equation:

$$F' U_L = \frac{(MC)\delta_w}{t_o A_t} \quad (15)$$

Using this heat loss factor, the optical efficiency is calculated using Eq. (13).

9. SOLAR BOX COOKERS

Box-type solar cookers/ovens use the greenhouse effect to cook food. A transparent glass or plastic cover over the insulated box allows short-wave radiation to pass through. The black-coated inner box absorbs this radiation. With the increase in the temperature of the inner tray, energy is reradiated at longer wavelengths. The glass cover does not allow long-wave radiation to pass through it. The temperature of the box increases and an equilibrium temperature is reached, where the input of solar energy is balanced by the heat losses. In order to enhance the performance of this cooker, a plane reflector is attached to the box. This type of cooker can hold a number of pots of food. Food does not burn, as it is cooked at low temperatures. Due to its light weight, it is easy to carry. The cost of this type of cooker is less than that of concentrating cookers.

The most popular box solar cooker was designed by M. K. Ghosh, in 1945. It consists of a rectangular aluminum tray. This tray is placed in a box made of fiberglass-reinforced epoxy resin, wood, aluminum, and a galvanized iron sheet. The space between the aluminum tray and the outer box is filled with insulation. The tray is covered with a movable double-glass cover, attached by a hinge to one side of the outer box. A flat glass mirror encased in a fiberglass-reinforced epoxy resin shell is fixed to serve as a reflector and also as a cover when the cooker is closed. A hinge arrangement is provided to tilt the reflector at different angles to focus solar radiation into the tray. Cylindrical tight-fitting vessels with flat bottoms and top lids are used for cooking food.

The aluminum tray, the outside of the cooking vessels, and the lids of the vessels are painted dull black. A temperature of 120°C can be attained in this type of cooker, which is sufficient for boiling-type cooking. Keeping the general configuration intact, a number of alternate materials have been used by different designers to reduce costs and to use local materials. Aluminized Mylar film/anodized aluminum sheets have been used for the reflector. Agricultural residues, crushed newspapers, and other materials have been used for insulation. Cardboard has been used for the outer box and top lid. Plastic sheets have been used in place of glass. Although these cookers are cheap, food can be cooked properly in these configurations at only climatically favorable places. The lifetime of these cookers is shorter than that of custom-made cookers. See Fig. 16.

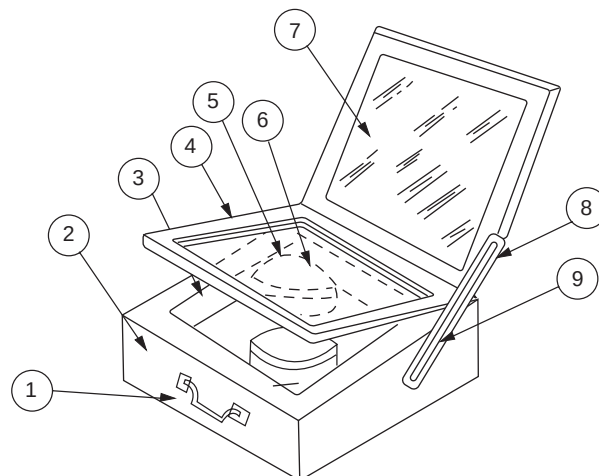


FIGURE 16 Box-type solar cooker. (1) Handle; (2) outer box; (3) insulation material (glass wall); (4) cover; (5) glass sheet; (6) cooking container; (7) plane mirror; (8) mirror support; and (9) hinged adjuster and guide.

9.1 Specifications of the Components (Bureau of Indian Standards)

9.1.1 Cover Plate

The inner cover plate is constructed of toughened glass 3 mm in thickness; the outer cover plate is constructed of plane glass 4 mm in thickness. Spacing is 10 × 2 mm. Solar transmittance should be 80% at a minimum. The aperture area is 500 × 500 mm. Tolerance is ± 5 mm.

9.1.2 Cooking Tray

The size should be $500 \pm 5 \text{ mm} \times 500 \pm 5 \text{ mm} \times 82 \pm 2 \text{ mm}$. The construction material is aluminum. The sheet should be $0.56 \pm 0.05 \text{ mm}$. The tray should be painted in a matte black finish to withstand temperatures up to 175°C. A scratch test should be carried out on a sample 10 × 10 mm in area.

9.1.3 Cooker Box

The material thickness for the galvanized iron sheet should be $48 \pm 0.05 \text{ mm}$ and that for the aluminum sheet should be $56 \pm 0.05 \text{ mm}$. The FRP should be 2 mm, at a minimum. The size of the cooker box is $600 \pm 5 \times 600 \pm 5 \times 500 \pm 3 \text{ mm}$.

9.1.4 Gasket

The gasket should be constructed of neoprene or silicon rubber and have a thickness of 3 mm.

9.1.5 Insulation

Insulation should consist of glass wool or mineral wool in 50 mm thick uniform pads. The insulation should be free of volatile material at 200°C.

9.1.6 Reflecting Mirror

The reflecting mirror should be $540 \pm 5 \text{ mm} \times 540 \pm 5 \text{ mm}$ in size. The thickness should be 3 mm. Solar reflection should be equal to 65%.

9.1.7 Cooking pots

The construction material for the cooking pots should be aluminum or stainless steel. The depth of the pots should be $63 \pm 2 \text{ mm}$. For aluminum pots of $200 \pm 5 \text{ mm}$ or $150 \pm 5 \text{ mm}$ in diameter, the thickness should be 1 mm. There should be three pots. For stainless steel pots of $200 \pm 5 \text{ mm}$ or $150 \pm 5 \text{ mm}$ in diameter, the thickness of the pots should be 50 mm. Two pots are needed.

9.1.8 Wheels

The wheels should be 25 mm in diameter, with ball bearings.

9.2 Performance Evaluation of Box-Type Cookers

Energy requirements are at a maximum during the heating period in boiling-type cooking and only a small amount of heat is required for the physical and chemical changes involved in the cooking process. Once the cooking temperature is attained, energy is required only to compensate for thermal losses and that required for physical and chemical changes and so the cooking is practically independent of the heat input rate. Thus, the time required to cook the same food in different cookers will depend on the heating period.

Two parameters have been recommended to evaluate the performance of box-type solar cookers. These are the first figure of merit, F_1 , and the second figure of merit, F_2 . The first figure of merit depicts optical efficiency and is the ratio of optical efficiency to heat loss factor; it is evaluated by a stagnation test without load. The second figure of merit, F_2 , gives an indication of the heat-exchange efficiency factor (F') and involves heating of a full load of water.

In a box-type solar cooker, there is no control over its temperature and the operation is transient, in contrast to a solar water heating system, where a steady state is obtained by circulating water at different temperatures through the tubes. A quasi-steady state is achieved when a stagnation temperature is obtained. The heat loss factor at stagnation, U_{LS} , can be calculated.

Only cookers with good optical transmission of glass covers and a low overall heat loss factor will

have a high value for the first figure of merit (F_1). However, good heat transfer to the vessel and its contents will be ensured by the second figure of merit, F_2 . Since there is no arrangement for the flow of liquid to obtain a steady-state withdrawal of energy from the cooker, as is the case with solar water heaters, it is difficult to measure the heat-exchange efficiency factor F' between the cooker plate and the contents of the pot. A factor, F_2 , is proposed for this purpose; it includes F' as well as U_{LS} and CR for the cooker utensil combination.

During both tests, the mirror booster is not used and the measurements are started approximately 2 h before solar noon.

To estimate the first figure of merit, F_1 , a solar cooker with utensils is kept in the sunshine and the rise in plate temperature, the solar radiation intensity, and the plate temperature are recorded. When the stagnation temperature is reached, the ambient temperature and solar insulation are measured. The figure of merit, F_1 , is calculated by

$$F_1 = \frac{\eta_o}{U_{LS}} = \frac{T_{ps} - T_a}{I_{Gs}}, \quad (16)$$

where η_o is the optical efficiency; I_{Gs} is the insulation on the horizontal surface at the time the stagnation temperature is reached (in watts per square meter); U_{LS} is the heat loss factor at stagnation; T_{ps} is the stagnation plate temperature (in degrees centigrade); and T_a is the plate temperature (in degrees centigrade).

To evaluate the second figure of merit, F_2 , the solar cooker without the reflector but with a full load of water and utensils is kept in the sun. Initial water is kept at an average temperature between ambient and the boiling point. Temperature and solar radiation are measured until the temperature reaches 95°C .

The second figure of merit, F_2 , is calculated using the equation

$$F_2 = \frac{F_1(MC)_w}{A(t_2 - t_1)} \times \ln \left\{ \frac{1 - (1/F_1)[(T_{w2} - T_a)/I_G]}{1 - (1/F_1)[(T_{w1} - T_a)/I_G]} \right\}, \quad (17)$$

where F_1 is the first figure of merit from the stagnation test; $(MC)_w$ is the product of the mass of water and specific heat (in joules per degrees centigrade); A is the aperture area of the cooker of the cover plate (in square meters); $t_2 - t_1$ is the time taken for heating from T_{w1} and T_{w2} (in seconds); T_a is the average air temperature during the time period $t_2 - t_1$ (in watts per square meter); and I_G is the radiation during the time period $t_2 - t_1$ (in watts per square meter).

It has been assumed that the global radiation, I_G , remains constant during the test. This assumption introduces some error into the data.

9.3 Testing Methods for Box-Type Solar Cookers (Bureau of Indian Standards)

The following tests have been recommended by the Bureau of Indian Standards (BIS) for evaluating the durability and thermal performance of box-type solar cookers. These tests include leakage tests, breakage tests, reflectivity tests, exposure tests, and thermal performance tests.

9.3.1 Leakage Tests

9.3.1.1 Cooking Tray Leakage Test

1. The cooking tray should be dismantled from the main body.
2. The cooking tray is filled with water; after 1 h, the joints of the tray should be examined for any signs of leakage.

9.3.1.2 Rubber Gasket Leakage Test

1. A piece of paper is inserted between the gasket and the cover plate in at least four positions along each side of the cooker.
2. The paper used should be 50 mm wide and 0.01 mm thick.
3. The cover plate should be properly tightened.
4. The paper should exhibit a firm resistance to withdrawal by hand at all points tested.

9.3.1.3 Cover Plate Leakage Test Leakage from the cover plate may occur from the upper and lower sides. Therefore, the cooker should be tested using the following two tests:

i. Leakage test for the upper side of the cover plate:

1. The cover plate should be properly tightened.
2. A thin film of water should be poured onto the cover plate.
3. After 1 h, the cover plate should be examined for any signs of water having entered between the two glass sheets.

ii. Leakage test for the lower side of the cover plate:

1. The cooking pots are filled with water and kept in the cooking tray.

2. The cover plate is tightened.
3. The cooker is placed out in the open for exposure to the sun's rays for 4–5 h.
4. Then, the cooker should be placed in the shade for 15 min to allow any vapor to condense.
5. The cover plate is examined for any signs of water vapor entry between the inside and outside of the cover plate.

9.3.2 Rain Penetration Test

9.3.2.1 Apparatus The basic apparatus is a 5 mm spray nozzle and a balance.

9.3.2.2 Procedure

1. The closed cooker is sprayed with 10 liters of water on all sides using a spray nozzle at a pressure of 0.1 MPa.
2. Spray nozzles are directed downward from the cooker top and also toward the four corners of the cooker.
3. The cooker should be weighed before the test.
4. After the test, the external surfaces of the cooker are wiped dry and the cooker should be reweighed.
5. During wiping, transport, and placement on the weighing machine, the cooker should remain in a horizontal position.
6. The balance used to weigh the cooker should have an accuracy of at least 20 g.
7. The permissible gain in weight should not be more than 30 g.

9.3.3 Slam Tests

This test is performed to ensure that the mirror or cover plate will not be damaged when allowed to fall from the fully opened position as given below:

9.3.3.1 Cover Plate Slam Test With the lid open, the cover plate is lifted as high as possible and then it is allowed to fall to a closed position. There should be no damage to the glass sheets.

9.3.3.2 Slam Test Mirror The lid should be held at near vertical and is allowed to fall to a closed position. There should be no damage to the mirror, the cover plate, or any other part of the cooker.

9.3.4 Mirror Reflectivity Test

9.3.4.1 Apparatus The basic apparatus consists of a photovoltaic-based solarimeter and a stand for mounting a sample piece of the reflecting mirror. The

stand should have a vertical pointer of 15 cm in height and an arrangement for mounting the solarimeter in a plane parallel to the plane of stand and at a height of 30 cm.

9.3.4.2 Procedure

1. The stand is adjusted for normal incidence such that there is no shadow of the vertical pointer.

2. This sample of reflecting mirror (540 mm × 540 mm) is placed on the stand in an open space. The open space should be free of any shadows and it should not receive any reflecting radiation from the surroundings.

3. The solarimeter should be placed in the stand provided for it in such a way that the sensor is near normal ($\pm 10^\circ$) to the sun's rays. The solarimeter (R_1) is then read.

4. The position of the solarimeter is changed in such a way that the sensor now faces the mirror at a near normal angle. The solarimeter (R_2) is then read.

5. The reflectivity of the mirror is calculated as follows: $R = R_2/R_1$.

6. The test should be repeated six times at 10 min intervals.

7. The average of the six values is calculated. This value will represent the reflectivity of the mirror.

8. The permitted conditions of solar radiation during the test are that it should be greater than 600 W/m².

9.3.5 Exposure Test Apparatus

The basic apparatus consists of a solar pyranometer, along with recording devices.

9.3.5.1 Procedure

1. The solar cooker is left to stagnate, which may lead to the following types of possible degradation:

- Breakdown of rubber or plastic material;
- Outgassing from the insulating material;
- Discoloration or peeling of black paint on the cooking pots and cooking tray; or
- Deposition of water vapor, dust, or any other material inside the double-glass lid.

2. The solar cooker is left in an unshaded area for at least 30 days having irradiation levels of at least 14 MJ/m² on the horizontal surface.

3. These days need not be consecutive.

4. The cooking pots inside the cooker should be empty.

5. The mirror is placed vertically and the cooker should be oriented to face south.

9.3.6 Cooking Tray Paint Peeling Test

A 10 mm × 10 mm area of painted surface is selected from squares of 1 mm × 1 mm by scratching horizontal and vertical lines with a pointed tool. The surface portion is covered with adhesive cellophane tape and then the tape is removed with a jerk. The test surface should be inspected and there should be no peeling of paint.

9.3.7 Transmittance Test for Cover Plate

The solar radiation is measured with a pyranometer/photovoltaic-based solarimeter directly (without the glass) and then measured with the cover plate above the pyranometer and solarimeter. The ratio of the two readings should be at least 0.8.

9.3.8 Thermal Shock Test for Gasket/Sealant

The gasket/sealant of 10 cm in length should be kept in an electric oven at a temperature of 150°C for 4 h. It should be cooled in air for 2 h, again placed in the oven at 150°C for 4 h, and then again cooled in air. After the test, the gasket and sealant should exhibit no cracking or brittleness.

9.4 Thermal Performance Test

The thermal performance test should be conducted under the conditions given below; values of F_1 and F_2 are then calculated.

9.4.1 Testing of Box-Type Cookers

9.4.1.1 Stagnation Temperature Condition Apparatus The apparatus consists of a pyranometer and a platinum resistance thermometer along with a recording device.

9.4.1.2 Procedure The test for evaluating F_1 should be carried out as follows in the morning hours before 10:00 A.M. so that the stagnation temperature is achieved near solar noon.

1. The solar cooker is placed in full sun.
2. The reflector of the solar cooker is covered with a black cloth.
3. The temperature inside the solar cooker is measured using RTD for a continuous 5 min interval.
4. When the cooker temperature has reached a steady state, the final temperature inside the cooker (T_{ps}), the corresponding ambient air temperature (T_a), and solar insolation are noted.

The steady-state conditions are defined as a 10 min period when:

1. The cooker temperature is constant within 1°C;

2. Solar radiation is constant within 20 W/m^2 ;
3. Ambient temperature is constant within 0.2°C ; and
4. Solar radiation is greater than 600 W/m^2 .

The performance factor F_1 is calculated using Eq. (16).

9.4.2 Heat-Up Condition

The figure of merit F_2 should be estimated as follows.

9.4.2.1 Apparatus The apparatus consists of a pyranometer and a platinum resistance thermometer, along with a recording device.

9.4.2.2 Procedure The empty cooking pots are weighed and then filled with water at 50°F (15°C). They are reweighed and the mass of water is calculated. The pots are placed in the cooker from which the mirror has been removed or covered. A temperature probe of RTD is placed in the largest of the cooking pots with the measuring tip submerged in the water. The temperature probe lead should be sealed where it leaves the cooking pots and the cooker.

The ambient temperature and wind speed are measured throughout the test. The test should start in the morning between 11:00 and 11:30 A.M. of local solar time. If radiation and temperature are measured by spot checks, these should be no more than 5 min apart.

Constant monitoring at 30 s intervals or less is desirable, with averages of radiation recorded over 2 min intervals. The following measurements should be taken:

1. Water temperature is measured along with the exact time that that measurement was recorded.
2. The data recording is continued until the water temperature exceeds 95°C .
3. Initial and final temperature/time data pairs are chosen. The initial temperature should be between 60 and 65°C and the final temperature should be between 90 and 95°C . These are denoted T_{w1} , and T_{w2} , respectively, and the corresponding times are t_1 and t_2 , respectively.
4. The average air temperature (T_a) between times t_1 and t_2 is calculated.
5. The radiation recorded between the two points should not have varied by more than $\pm 10 \text{ W/m}^2$ and should always have exceeded 600 W/m^2 .
6. The average radiation (I_g) over the time t_1 and t_2 is calculated.
7. The second figure of merit, F_2 , is calculated as per Eq. (17).

8. As per the BIS standard, the value of F_1 should be >0.12 for grade A cookers and >0.11 for grade B cookers.

Theoretical analysis as well as experimental evaluation shows that it not possible to achieve these values with the BIS-proposed specifications. However, these values can be achieved by increasing optical efficiency and decreasing the overall heat loss coefficient.

10. MULTIPLE GLASS BOX COOKERS

It is difficult to cook food in a single reflector box-type solar cooker at higher latitudes during the winter season due to higher losses resulting from low ambient temperatures. To overcome this problem, folding two-step asymmetrical reflector box solar cookers have been designed. These consist of a hot box and two plane mirrors fixed at an angle of 13.5° . The hot box has double glazing at the top and insulation on the bottom and sides. At the end of cooking, mirrors can be folded onto the top of the hot box. See Figs 17 and 18.

Solar radiation through the aperture, AL_1 , AL_2 , and AL_3 , can be calculated by the following equations:

$$AL_1 = B \sin(\phi). \quad (18)$$

$$AL_2 = B \sin(\theta_1 - \phi). \quad (19)$$

$$AL_3 = B \sin(\theta_2 - \phi). \quad (20)$$

$$\theta_1 = \pi/3 + 2/3\phi. \quad (21)$$

$$\theta_2 = \theta_1 - 13.5. \quad (22)$$

See Fig. 19.

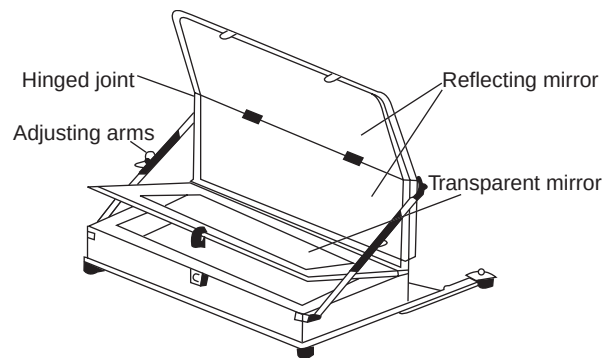


FIGURE 17 Multiple glass box cooker.

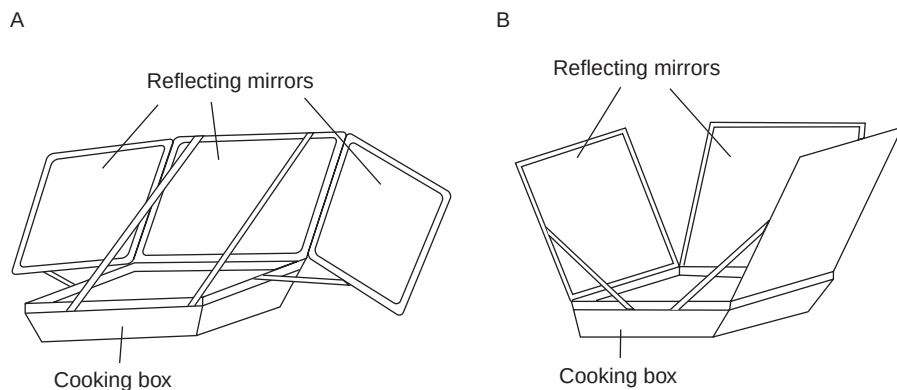


FIGURE 18 Multiple glass box cooker.

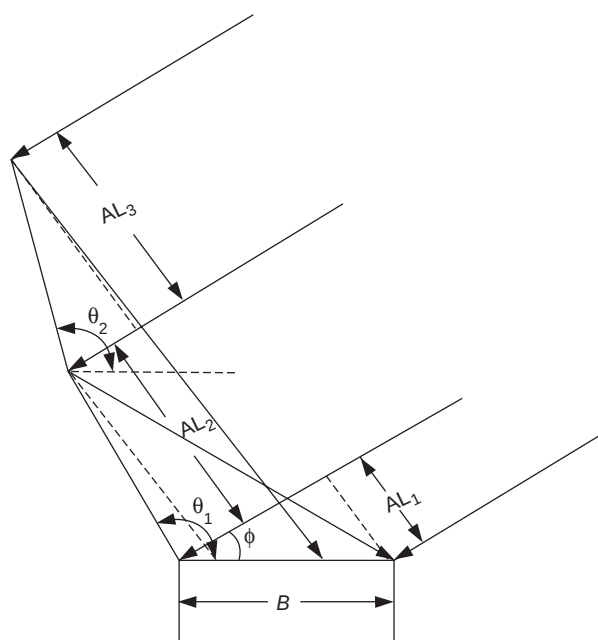


FIGURE 19 Depiction of the geometry used to determine the solar radiation through the aperture in multiple glass box cookers.

11. INDIRECT SOLAR COOKERS

There are no provisions in the conventional box-type or concentrator-type cookers for cooking inside the kitchen or during the period when solar radiation is not available. A number of strategies have been used for special types of indirect solar cookers.

11.1 Organic Fluid Cookers

In this type of cooker, organic fluid is heated using a cylindrical parabolic reflector/CPC/flat-plate collector outside the kitchen. The hot fluid is stored inside the kitchen using the thermosphon principle for cooking applications.

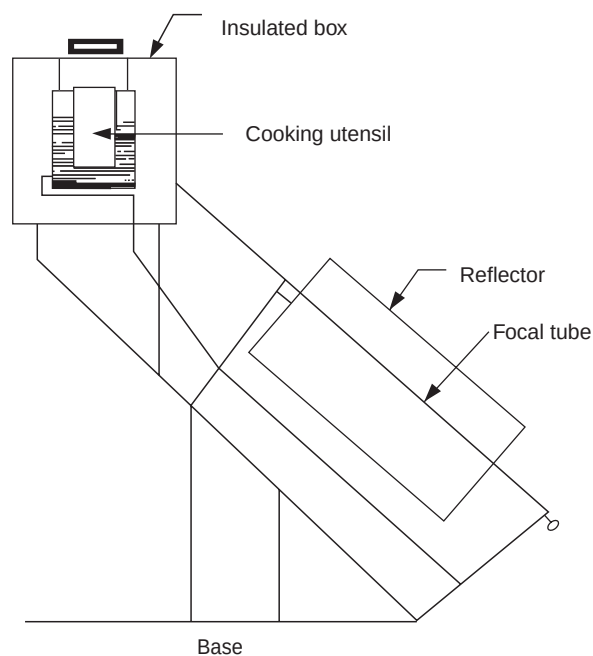


FIGURE 20 A typical configuration of an organic fluid cooker.

A typical configuration of this type of cooker is shown in Fig. 20.

11.2 Steam Cookers

This type of cooker is similar to the organic fluid cooker, except that low-pressure steam is used as the working fluid. Using flat-plate or concentrating reflectors, low-pressure steam is produced. Food is cooked in a double-jacketed vessel.

11.3 Chemical Cookers

These cookers are based on a reversible chemical reaction between a working fluid (gas or vapors) and



FIGURE 21 A typical configuration of a Scheffler community cooker.

a chemical compound. The reaction is endothermic in one direction and exothermic in the reverse direction. Working fluid becomes desorbed by solar energy and is stored for later use. When heat is required, the desorbed fluid is again allowed to adsorb in the chemical. The liberated heat is used for cooking.

12. SCHEFFLER COMMUNITY COOKER

This cooker has a 6.5 m^2 parabolic reflector dish, which is placed nearly 3 m away from the north-facing wall. The reflector reflects and concentrates solar radiation onto a secondary reflector through an opening in the wall. The secondary reflector deflects the radiation to the bottom of the cooking vessel. Temperatures of up to 400°C can be attained. Cooking can be carried out in a conventional manner in the kitchen. The reflector is tracked with a clockwork mechanism to ensure continuous cooking. There is an arrangement for seasonal adjustment of the reflector. Figure 21 shows the configuration of the cooker.

13. RATING OF SOLAR COOKERS

At the Third International Conference on Solar Cooker Use and Technology, the question of rating solar cookers based on field tests for comparing different models and types of solar cookers was discussed. It was proposed that apart from thermal performance, there are other important parameters that need to be considered from the user's point of view. The selection process in the field should take into consideration the particulars of the sites, conditions, and consumers. The rating criteria should include, in addition to thermal performance, critical

factors such as the following: (1) cost, (2) safety, (3) heating and cooking capacity, (4) convenience, (5) durability, (6) ease of maintenance, (7) stability in the wind, and (8) operating instructions on the use of the cooker.

For wider acceptability of the solar cooker, the cost should be reasonable so that ordinary users in developing countries can afford it. It should meet the cooking needs of an average family. It should be convenient to cook food in the cooker, as women do not prefer a gadget that is cumbersome to use. The equipment should be safe to work with and should not cause burn injuries while one is cooking or removing cooked food from the apparatus. As solar cooking is done outdoors, the construction material should be durable so that the cooker may last for a minimum of 8–10 years. Cookers should be easy to maintain as there is a lack of availability of technically trained personnel for maintenance in developing countries and there is a high cost of maintenance in developed countries. A maintenance manual should be provided. Under windy conditions, the cooker can be blown off or fall, resulting in breakage of the glass components and spoiling of the food. There should be provisions for holding the cooker steady under windy conditions. A complete operating manual should be provided in written text as well as in pictorial form to overcome language barriers.

SEE ALSO THE FOLLOWING ARTICLES

Solar Cells • Solar Cooling, Dehumidification, and Air-Conditioning • Solar Detoxification and Disinfection • Solar Distillation and Drying • Solar Energy, History of • Solar Fuels and Materials • Solar Heat Pumps • Solar Ponds • Solar Thermal Energy, Industrial Heat Applications • Solar Thermal Power Generation • Women and Energy: Issues in Developing Nations • Wood in Household Energy Use

Further Reading

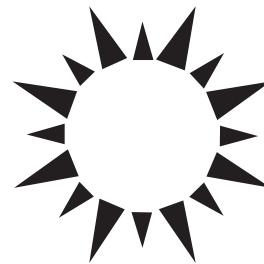
- Blum, B. (ed.). (1989). "The Solar Box Cooker Handbook." Solar Box Cooker International, Sacramento, CA.
- Bureau of Indian Standards (1992). Indian Standard No. IS 13429, Solar Cooker, Parts 1, 2, and 3.
- Duffie, J. A., Log, G. I. G., and Back, B. (1953). Laboratory and field studies in plastic reflector solar cookers. In "Proceedings of the United Nations Conference on New Sources of Energy E 35-587," Vol. 5, pp. 339–346.
- Garg, H. P., Bandyopadhyay, and Datta, G. (1985). Mathematical modelling of the performance of a solar cooker. *Appl. Energy* 14, 233–239.

- Garg, H. P., and Kandpal, T. C. (1999). "Laboratory Manual of Solar Thermal Experimentation." Narora.
- Ghai, M. L. (1953). Solar heat for cooking. *J. Sci. Industr. Res.* **12A**, 117–124.
- Ghosh, M. K. (1945). Solar cooker. *Bull. Inst. Eng.* **14**, 1–7.
- Habeebullah, M. B., Khalifa, A. M., and Olwi, I. (1995). The oven receiver: An approach toward the revival of concentrating solar cookers. *Solar Energy* **54**, 227–237.
- Kundapur, A. (1998). Review of solar cooker designs. *TIDE* **8**.
- Lof, G. O. G. (1963). Recent investigation in the use of solar energy for cooking. *Solar Energy* **7**, 125–132.
- Mullick, S. C., Khandpal, T. C., and Saxena, A. K. (1987). Thermal test procedure for box-type solar cookers. *Solar Energy* **39**, 353–360.
- Shukla, J. K., and Patel R. C. (1990). Study of connective heat losses in solar cookers. In "Proceedings of the National Solar Energy Convention" (Udaipur, A. N. Mathur, and N. S. Rathore, Eds.). Himanshu Publications.

Solar Cooling, Dehumidification, and Air-Conditioning

GERSHON GROSSMAN

Technion—Israel Institute of Technology
Haifa, Israel



-
1. Introduction to Solar Cooling
 2. Thermodynamic Principles
 3. Closed-Cycle Systems
 4. Open-Cycle Systems
 5. Conclusion

Glossary

absorption A thermodynamic cycle powered by heat and employing a working fluid pair (absorbent/refrigerant) to produce cooling.

closed-cycle system Cooling/dehumidification device in which the working fluid is enclosed in the machine and interacts with the cooled medium by heat transfer through solid walls.

coefficient of performance (COP) Ratio of the cooling produced to the heat/work supplied.

desiccant Hygroscopic material serving to attract and retain humidity from surrounding air.

heat engine Cyclic device receiving heat from a high-temperature heat reservoir, rejecting part of it to a low-temperature heat reservoir, and converting the difference to work.

heat pump Cyclic device extracting heat from a low-temperature heat reservoir and rejecting it to a high-temperature heat reservoir, powered by work.

heat reservoir A source/sink of heat at constant temperature.

open-cycle system Cooling/dehumidification device in which the working fluid (water and/or air) forms part of the cooled medium.

The possibility of providing cooling and air-conditioning by means of energy from the sun has attracted the attention of humans since the early development of solar technology. The necessity of air-conditioning for thermal comfort in hot areas of

the world and the abundance of sunshine in these areas have intrigued researchers with regard to how to combine the two for people's benefit. Furthermore, in contrast to other solar applications such as heating, the greatest demand for cooling occurs when the solar radiation is most intense, thus making its use for this application all the more attractive.

1. INTRODUCTION TO SOLAR COOLING

Among the various thermal applications of solar energy, cooling is one of the more complex, both in concept and in construction. This is no doubt one of the reasons why its utilization is not as widespread as space or water heating. It is not sufficient to collect, store, and distribute the solar heat. The energy must be converted to cold by means of a heat-actuated device capable of extracting heat at a low temperature from the conditioned space and rejecting it into the higher temperature of the outside air. In thermodynamic terms, a heat-pumping operation is required.

A schematic diagram of a complete solar air-conditioning system is presented in Fig. 1 showing the different components common to most such systems. A heat transfer fluid is heated in the solar collectors to a temperature well above ambient and used to power a cooling device—a type of heat-actuated heat pump. The heat transfer fluid may be air, water, or another fluid; it can be stored in a hot state for use during times of no sunshine. Heat extracted by the cooling device from the conditioned space and from the solar energy source is rejected to the environment using ambient air or water from a cooling tower. The cooling device may produce the cooling effect in a number of different ways. It can

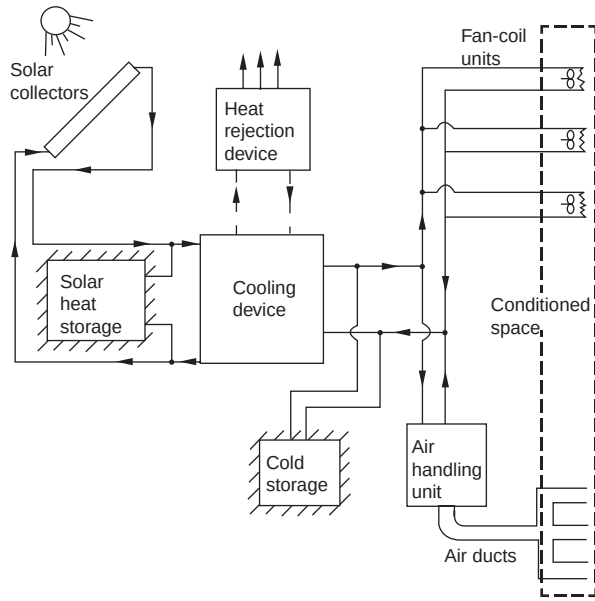


FIGURE 1 Schematic diagram of a solar cooling system.

generate chilled water that may be distributed through fan coil units or used to cool air, which is then distributed through air ducts; the chilled water may be stored for use when the cooling device is not in operation. Alternatively, the machine can cool the conditioned air directly.

Solar collectors are a major part of every solar cooling system. They convert solar radiation into heat at a temperature suitable for powering the cooling device. The collectors vary in type from low-temperature flat plates to more sophisticated high-temperature evacuated tubes or concentrators.

Growing demand for air-conditioning in recent years, particularly in hot and humid climates, has imposed a significant increase in demand for primary energy resources. Electric utilities experience peak loads on hot summer days and are often faced with brown-out situations, barely capable of meeting the demand. With suitable technology, solar cooling can help alleviate, if not eliminate, the problem. This article describes trends in solar cooling, dehumidification, and air-conditioning. It is divided into two main parts and considers closed-cycle and open-cycle alternatives, each with its own advantages and limitations.

2. THERMODYNAMIC PRINCIPLES

In considering the various options for heat pumping systems that may be employed for solar cooling, one

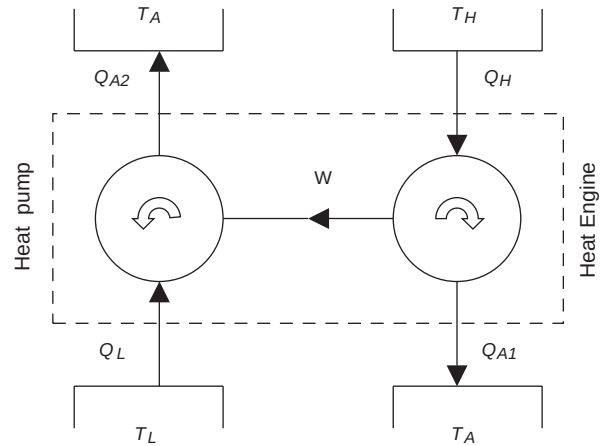


FIGURE 2 Thermodynamic description of a heat-powered cooling device.

must take into account the thermodynamic principles of a heat-powered cooling device. Typically, the system operates between heat reservoirs at three temperature levels: T_L , the low temperature of the cooled medium or the conditioned space; T_H , the high temperature supplied by the solar heat source; and T_A , the intermediate temperature of the environment serving as a sink for the reject heat. The cooling device may be described in basic thermodynamic terms as a combination of a heat engine and a heat pump, as illustrated in Fig. 2. The former receives heat Q_H from the high (solar) heat reservoir at T_H , rejects part of it to the heat sink at T_A , and converts the rest to mechanical work W . The latter employs this work to pump heat Q_L from the low-temperature reservoir at T_L to the heat sink at T_A . According to the second law of thermodynamics, best results are obtained from this machine when both cyclic devices are reversible and operating at the Carnot limit. Under these optimum conditions, the overall coefficient of performance (COP) of the cooling device is found to be

$$\text{COP} = \frac{Q_L}{Q_H} = \frac{T_L}{T_H} \frac{(T_H - T_A)}{(T_A - T_L)}. \quad (1)$$

As evident from Eq. (1), the performance of the heat-powered cooling device is highly dependent on the temperature of the heat source, with better results obtained with higher T_H . The type of solar collector employed to supply the heat thus imposes a limitation on the COP. With common flat-plate collectors, the heat is obtained at relatively low temperatures. Concentrating collectors reach higher temperatures but are more expensive and complex to operate due to their need to track the sun. Another conclusion

from Eq. (1) is that performance improves by lowering the heat rejection temperature T_A . From this standpoint, heat rejection by water cooling is preferred to that by air cooling.

3. CLOSED-CYCLE SYSTEMS

In practice, there need not be a clear distinction between the heat engine and the heat pump parts of the machine. Cycles such as absorption or ejector (jet) have been proposed in which the work produced by part of the machine is “invisible” and goes directly to power the other. Nevertheless, the fundamental thermodynamic principles of the cooling device described previously still hold and set the Carnot upper limit for the COP as given by Eq. (1).

During the great surge in solar energy research and development following the energy crisis of the 1970s, there have been several attempts to build solar-driven cooling systems by means of a solar heat engine driving a heat pump, according to Fig. 2. The former varied from Stirling engines to Rankine engines employing organic fluids; the latter were conventional vapor-compression cooling devices with common refrigerants. These efforts have generally not gone beyond the laboratory. Experience has shown that closed-cycle systems most suitable for

solar cooling are based on absorption cycles, which totally eliminate the mechanical interaction between the two parts of the machine. Absorption systems have many advantages: operation at wide ranges of heat source and heat sink temperatures; multistaging for improving COP with high-temperature heat sources; quiet operation with no moving parts; and environmentally safe working fluids. Moreover, absorption systems are amenable to combining available solar heat with backup heat for meeting the cooling demand.

Figure 3 is a schematic diagram of a typical absorption chiller in the simplest, single-effect configuration. The system employs the working fluid pair lithium bromide (LiBr)–water, in which the water serves as the refrigerant and the LiBr as the absorbent. The diagram is plotted in a system of coordinates showing the temperature on the horizontal axis and the pressure prevailing in the system's exchange units on the vertical axis.

The operation of single-effect absorption chillers is well documented in the literature and is described only briefly here. Liquid refrigerant (water) leaving the condenser at state 12 in Fig. 3 expands through the throttle valve (6) into the evaporator (1), where it evaporates from state 13 to 9, taking its heat of evaporation from the stream of chilled water and cooling it from state 1 to 2. The vapor leaving at

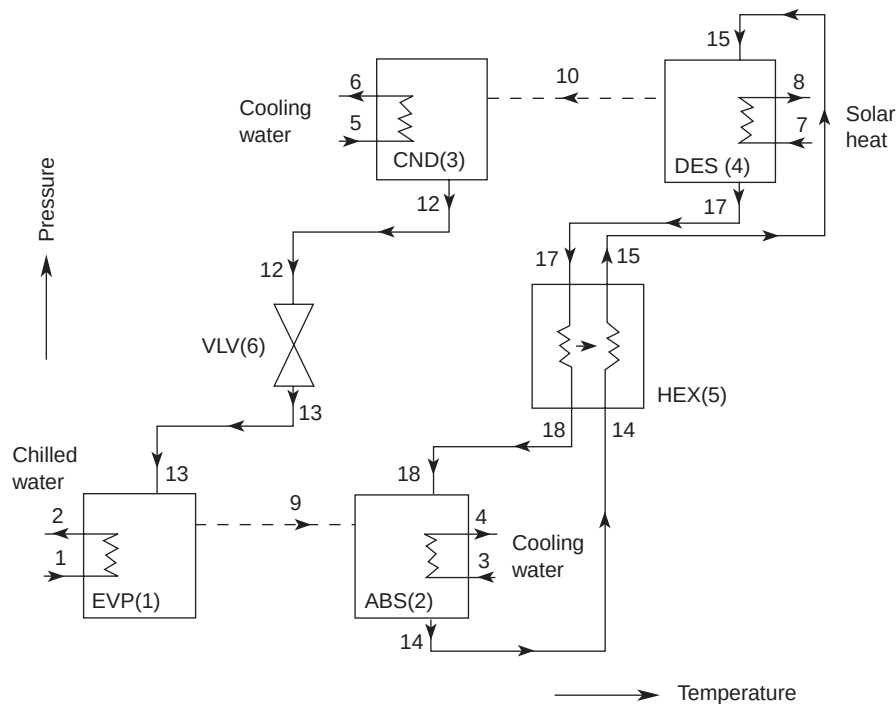


FIGURE 3 Schematic description of a single-effect LiBr–water absorption chiller.

state 9 is absorbed in the absorber (2) by a hygroscopic solution of LiBr entering concentrated (strong) at state 18 and leaving diluted (weak) at state 14. The exiting solution is pumped via heat exchanger (5) to desorber (4) by means of a small pump (not shown), requiring a small and negligible amount of work. The desorber serves to regenerate the solution from a weak to a strong state by applying to it heat from the solar-heated water stream 7–8, causing the desorption of water at state 10. The solution entering the desorber (4) at the weak state 15 is thus concentrated to state 17 and returns to the absorber (2) through the recuperative heat exchanger (5). The desorbed vapor at state 10 condenses in the condenser (3) to liquid state 12, then expands into the evaporator. The absorber and condenser are cooled by streams of cooling water 3–4 and 5–6 to reject the heats of absorption and condensation, respectively. As evident, the system operates between two pressure levels and interacts with heat sources/sinks at three temperature levels: low-temperature cooling in the evaporator, intermediate-temperature heat rejection in the absorber and condenser, and high-temperature (solar) heat supply in the desorber. These correspond to T_L , T_A , and T_H , respectively, in Fig. 2 and in the Carnot equivalent of Eq. (1).

From the previous description, it is clear that an absorption system employs a working fluid pair (absorbent/refrigerant), unlike the vapor compression heat pump, which employs a single fluid. The previously mentioned LiBr–water is the most common fluid pair in solar as well as other heat-powered absorption air-conditioning systems. Another fluid pair is water–ammonia, which generally requires higher firing temperatures and yields lower performance but can produce below-freezing temperatures and hence is suitable for refrigeration applications. Other working fluid pairs have been proposed but are rarely found in solar applications.

A great deal of experience is available on solar-powered absorption systems from demonstration projects carried out in the past three decades, most of which have employed single-effect LiBr–water chillers with low-temperature solar collectors. The first projects employed conventional machines originally designed for gas- or oil-fired operation. The heat to the desorber of these units was supplied by hot water from a solar heat source (e.g., flat plate, evacuated tube, or concentrating collectors). However, lowering the desorber temperature with respect to the original fuel-fired design—without making proper alterations in the machine—resulted in con-

siderable downgrading of the entire device. Attempts were made to redesign the units and modify the desorber in particular, such that they could operate efficiently with solar-heated water. The first commercial absorption air conditioner built specifically for solar operation was the 3-TR Solaire unit by Arkla Industries based on an earlier gas-fired model. This unit is well-known among solar cooling engineers because it was used in a large number of demonstration projects. Its cost, however, was quite high for its small size. Soon thereafter, other manufacturers brought to the market new and improved absorption chillers based on LiBr–water and designed specifically for solar operation.

Two basic requirements must be considered by the designer of a solar absorption system that distinguish it from conventional fuel-fired systems. The first is the need to operate at a low temperature of the heat source: The temperatures available from flat-plate collectors are usually limited to below 90°C. The other is the need for a backup system that can supply the cooling load when insufficient or no energy is available from the sun. For the former requirement, a desorber design based on falling film instead of the conventional pool type has proven effective. Also, a solution preheater placed between the heat exchanger outlet and the desorber inlet was found to be effective in reducing the size and cost of the desorber and improving performance.

The problem of backup is more complicated. A good design must not only supply the full cooling demand at times of low or no sunshine but also make use of whatever solar radiation is available and take only the rest from primary sources. One possible approach is to heat the desorber of the same chiller both from the solar collectors and from an additional source of hot water (e.g., a fuel-fired boiler). However, the hot water coming from the boiler will normally interfere with the relatively colder water available from the collectors and hence drastically reduce the use of available solar radiation. Simulation results show that under conditions of low sunshine, the collectors connected in series with the backup boiler can turn into a heat sink instead of a heat source. Another possibility is to operate a second cooling machine powered by primary energy alongside the solar machine powered only from the sun, with the former supplying whatever load the latter cannot deliver. This requires installation of a complete additional unit, which would be idle a significant portion of the time. To solve the problem effectively, it has been proposed to incorporate an auxiliary desorber in the single-effect absorption

cycle. The auxiliary desorber is powered by the backup, whereas the original desorber is solar powered. The weak solution from the absorber goes first to the solar-powered desorber, where it is concentrated as much as possible with the available solar heat; it then proceeds to the auxiliary desorber and is concentrated further by means of heat from the backup source. The vapor from both desorbers is sent to the condenser. Other variations of this cycle have been proposed.

Experience from the numerous solar-powered air-conditioning demonstration projects with single-effect LiBr–water absorption systems has shown their main limitation to be cost, dominated by the solar part of the system. Single-effect absorption systems are limited in COP to approximately 0.7 and hence require a large collector area to supply the solar heat needed for their operation. This area could be reduced by employing systems with improved COP, which may be achieved with a higher temperature of the heat source, as indicated by

Eq. (1). A single-stage absorption system is not suited to utilize a heat source at a temperature higher than approximately 100°C ; unlike other heat-activated heat pumps that follow the Carnot trend, the absorption COP levels off with the increase in desorber temperature due to increased mixing and circulation losses. To take advantage of a high-temperature heat source, absorption systems must be configured in stages. The principle is to utilize the heat rejected from the condenser to power additional desorbers, thereby approximately doubling or tripling the amount of refrigerant extracted out of solution with no extra solar heat spent.

The principle of multistaging may be explained by referring to Fig. 4, which depicts a double-stage LiBr–water absorption chiller comprising two desorbers and two condensers serving one absorber/evaporator combination. The system may be regarded as an extension of the single-stage chiller of Fig. 3, with an additional condenser (7), desorber (8), heat exchanger (9), and expansion valve (10). As

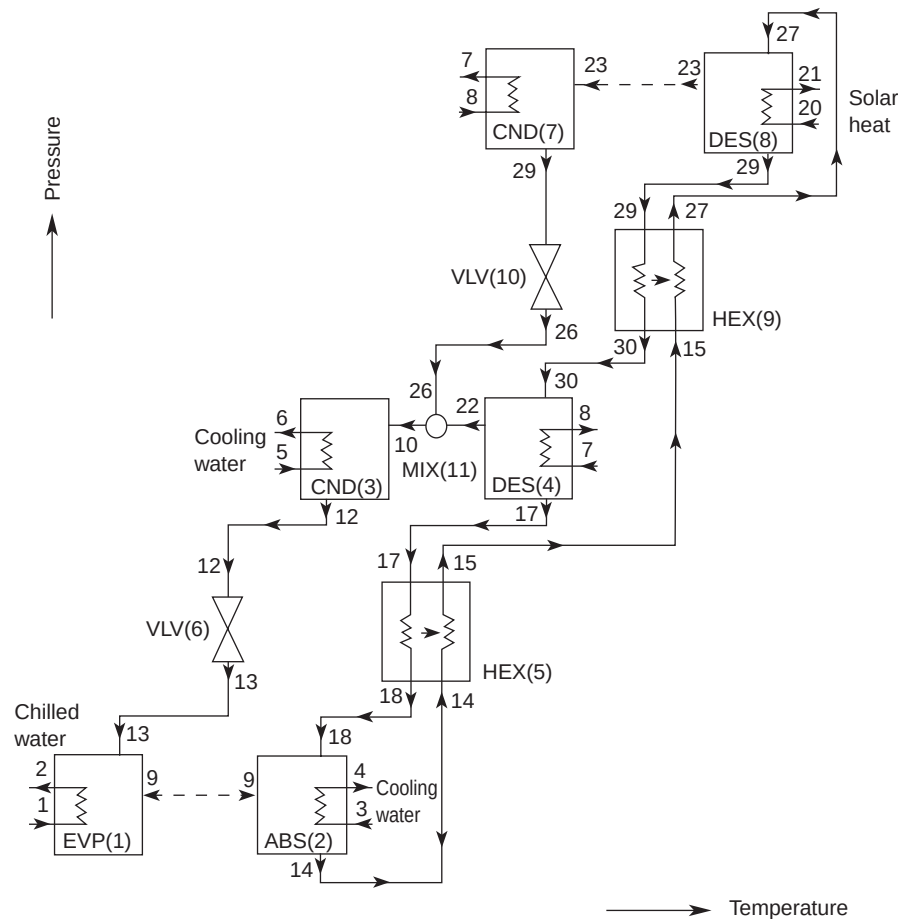


FIGURE 4 Schematic description of a double-effect LiBr–water absorption chiller.

evident, the system operates at three pressure levels and four temperature levels. The weak solution leaving the absorber (2) at state 14 is circulated to the two desorbers, connected in series, and regenerated in two stages, with each desorber doing part of the regeneration task. High-temperature solar heat is applied to the high-temperature desorber (8), whereas the low-temperature desorber (4) receives the reject heat from the high-temperature condenser (7) through the hidden loop 7–8. The condensate from both condensers expands into the evaporator and yields approximately double the cooling effect for the single-stage system of Fig. 3. The system is thus termed double-effect. The same principle may be extended with three desorbers and three condensers to give a triple-effect system when sufficiently high temperature of the heat source is available.

As mentioned previously, most solar-powered absorption cooling projects have utilized single-effect systems with low-temperature solar collectors. Due to developments in gas-fired absorption systems in recent years, mainly in the United States and Japan, higher COP LiBr–water chillers have become available: double-effect systems, available on the market, with COP in the range 1.0–1.2, and triple-effect systems, still under development but close to market, with estimated COP of approximately 1.7. These systems may be adapted to and employed in a solar-powered installation with high-temperature solar collectors. Figure 5 compares the performance of several multieffect chillers, showing the COP as a function of the solar heat supply temperature for typical single-, double-, and triple-effect chillers with the same size components under the same operating conditions: cooling water inlet at 30°C and chilled

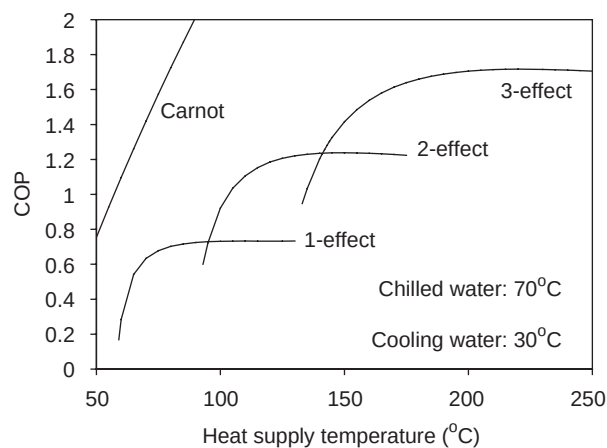


FIGURE 5 Coefficient of performance (COP) as a function of solar heat supply temperature for single-, double-, and triple-effect LiBr–water absorption chillers.

water outlet at 7°C. The corresponding Carnot performance curve, according to Eq. (1), is plotted for comparison. It is evident that for each absorption system, there is a minimum value of the heat supply temperature below which it does not perform at all. As this temperature increases, the COP increases sharply, then levels off to some asymptotic value. The reasons for this behavior are well understood and documented but are beyond the scope of this article. The single-effect system gives best results in the temperature range 80–100°C; above these temperatures, it is worthwhile to switch to the double-effect system, up to approximately 160°C, and then to the triple-effect system. Note, however, that with a double-effect system, if the solar-supplied temperature should drop below 100°C, the performance will drop sharply—below that of a single-effect system. It is therefore of utmost importance to incorporate a backup in a way that will ensure the supply of heat above the required minimum temperature.

With the availability of high-temperature gas-fired systems, the following question arises: Due to their high COP, are multieffect systems more suitable for solar-powered cooling than their single-effect counterparts? It is attempted to answer this question by conducting a comparison between single-, double-, and triple-effect systems. The results of the comparison clearly indicate that the economics of solar-powered air-conditioning is dominated by the cost of the solar part of the system. This seems to show preference for the multieffect absorption systems, whose higher COP results in less collector area per kilowatt of cooling. Despite this fact, the higher cost of the high-temperature collectors often makes the total system cost higher than that of the low-temperature systems. The challenge before the solar industry is to reduce the price of high-temperature collectors, which will pave the way to employ the high-COP absorption cooling machines.

4. OPEN-CYCLE SYSTEMS

Open-cycle desiccant-type cooling systems represent a potentially promising alternative to other methods of solar-powered air-conditioning. Although the main thrust in research on novel closed-cycle absorption systems has been toward increasing the operating temperature in order to improve efficiency through multistaging, open-cycle absorption and desiccant systems have been developed for use with low-temperature heat sources such as flat-plate solar collectors. Relatively little work has been done in this

area in comparison with studies on the high-temperature alternative, despite the high cost of the latter and its various limitations: the need for sun tracking with concentrating solar collectors, their ability to use only the direct component of the solar radiation, and difficulty in storing high-temperature heat.

Desiccant systems are essentially open absorption cycles, utilizing water as the refrigerant in direct contact with air. The desiccant (sorber) can be either solid or liquid and is used to facilitate the exchange of sensible and latent heat of the conditioned airstream. The term open is used to indicate that the refrigerant is discarded from the system after providing the cooling effect and new refrigerant is supplied in its place in an open-ended loop. In both liquid and solid desiccant types of systems, the process air is treated in a dehumidifier and goes through several additional stages before being supplied to the conditioned space. The sorbent is regenerated with ambient or exhaust air heated to the required temperature by the solar heat source.

The operation of the desiccant cycle is made possible by the property of desiccants of attracting and holding water vapor from the surrounding air at low temperatures and releasing water to air at high

temperatures. The exchange of water vapor between air and desiccant depends on the relative magnitudes of water vapor pressure in the air and on the surface of the desiccant, respectively. The vapor pressure in the air is a function of the water mole fraction, per Dalton's law, and independent of temperature. The vapor pressure in the desiccant is a function of its temperature and water content: Vapor pressure increases with both temperature and water content. Thus, at low desiccant temperatures dehumidification of air will take place in a sorption process, whereas at high desiccant temperatures the same air could be used to regenerate the desiccant—that is, reduce its water content in a desorption process. A sorption process is always exothermic; that is, it generates heat due mainly to the condensation of vapor on the surface of the desiccant as well as to the heat of mixing (in the case of liquids) or heat of wetting (in the case of solids). Conversely, desorption is endothermic; that is, it requires heat input to separate and evaporate water from the desiccant.

A desiccant air-conditioning system in its simplest form consists of conditioning and regenerating cycles, as illustrated in Fig. 6. The conditioning cycle consists of a dehumidifying stage with simultaneous or subsequent removal of the heat of sorption. Warm

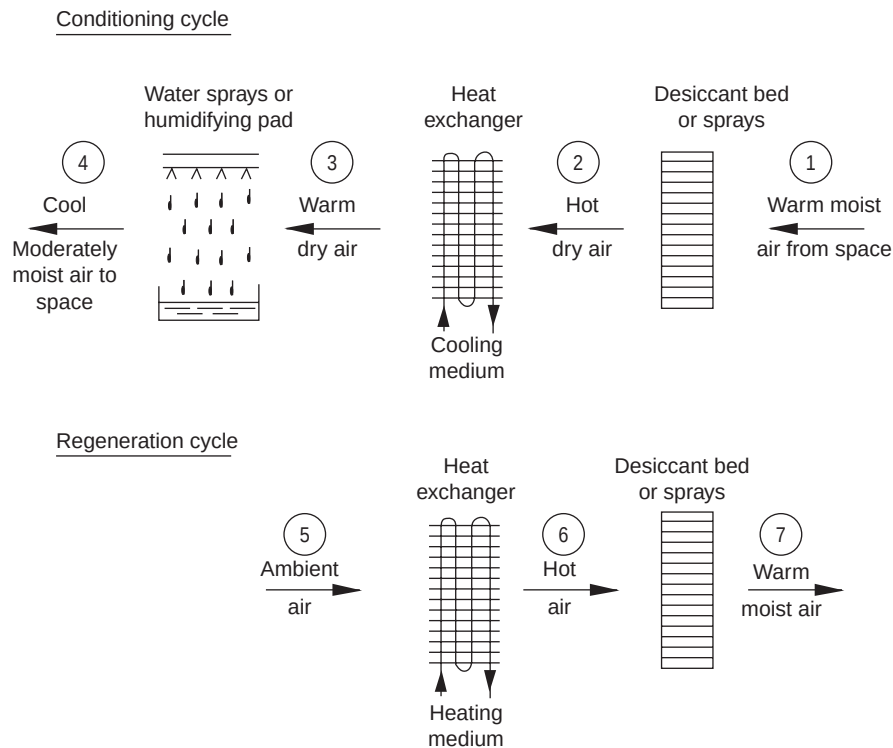


FIGURE 6 Principle of open-cycle desiccant air-conditioning.

and moist process air from the conditioned space or from the outside (state 1) is first brought into contact with an active desiccant. Water vapor is removed from the air and absorbed (in a liquid desiccant) or adsorbed (on a solid desiccant). This exothermic process causes an increase in the temperature of the air, whereas its humidity is reduced (state 2). The hot and dry air is then cooled to near ambient temperature, without change in water content (state 3), in a heat exchanger using a moderately cool medium (cooling tower water, exhaust air, or evaporatively cooled ambient air). If necessary, a humidifying stage 3–4 is added in which evaporative cooling is carried out. Sensible heat is changed into latent heat; that is, the air temperature is reduced while its water content is increased to slightly below the original value at state 1. This air thus has both sensible and latent cooling capabilities when released into the conditioned space.

The capacity of the desiccant to attract water vapor decreases with the increase in its water content, and at a certain point it has to be regenerated. In the regeneration cycle illustrated in Fig. 6, ambient air at state 5 is heated in a heat exchanger with solar-heated water (or equivalent) to state 6 and is brought into contact with the desiccant. Water vapor is transferred from the desiccant to the air in the desorption process 6–7. Warm and humid air at state 7 is exhausted and the dry desiccant is ready to be used again in the conditioning cycle.

Desiccant systems possess several advantages relative to their closed-cycle counterparts: They operate at ambient pressure and not in a vacuum or at an elevated pressure; heat and mass transfer between the air and the desiccant take place in direct contact; and both cooling and dehumidification of the conditioned air may be provided, in variable quantities, to fit the load in the conditioned space. The latent load is handled without the necessity to cool the air below its dew point; consequently, reheat is never required. The following are disadvantages: They have a low COP due to inherently inefficient regeneration; relatively large air volumes must be pumped, leading to potentially high parasitic losses; and contamination of the desiccant by dirt and dust contained in the air may require its replacement after some period of operation.

Solid desiccant systems employ adsorbents (rather than absorbents) such as silica gel or various types of molecular sieves. The humidity is adsorbed on the surface of the desiccant, which is highly porous. Since the solid desiccant cannot be circulated by pumping, these systems usually employ a rotary bed

carrying the sorbent material, referred to as a desiccant wheel. The wheel features continuous operation of the system by rotating between the process and regenerating airstream. A typical desiccant wheel is shown in Fig. 7. It consists of a rotary bed or drum made of an inert material, corrugated or extruded, forming axial flow channels for the air. The wheel is packed or impregnated with the desiccant material, covering the walls of the channels. The wheel in Fig. 7 forms part of a natural gas-boosted solar air-conditioning system, employing molecular sieves as desiccant. The wheel is divided into four sections: the sorption section, through which the process airstream flows; the low-stage regeneration section, in which desorption of the desiccant by means of solar-heated air takes place; the high-stage regeneration section, in which desorption is boosted by flue gas from a burner; and a purge section in which cooling of the matrix, as it leaves the regeneration section, is carried out. The purge section ensures that the desiccant entering the process airstream is cooled sufficiently for dehumidification to take place, thus reducing the average process airstream humidity and temperature.

Solid desiccants usually allow a higher degree of dehumidification than the liquids but require a higher temperature for their regeneration. They usually operate according to the cycle depicted in Fig. 6. The past decade witnessed a significant increase in solid desiccant systems on the market, usually for gas-fired operation in applications requiring dehumidification in addition to conventional cooling, such as in supermarkets. Solar-powered operation with solid desiccants normally requires higher temperatures than required for liquids,

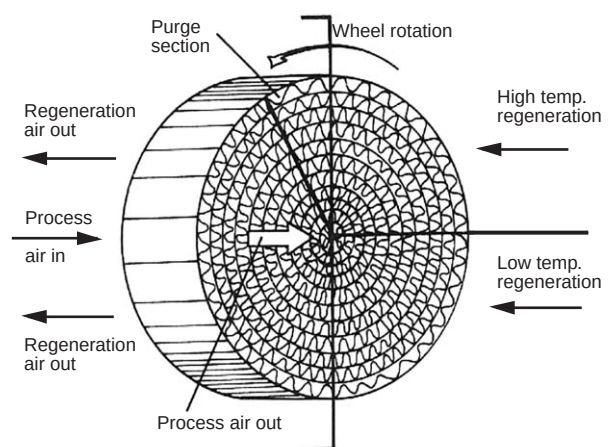


FIGURE 7 Diagram of a typical desiccant wheel.

possibly beyond the range of flat-plate collectors. The regeneration air may be heated directly in air-heating collectors or indirectly using water as a heat-carrying medium.

Liquid absorbents of water comprising aqueous solutions of lithium bromide, lithium chloride, or calcium chloride are employed in the open systems using a cycle similar to that of a closed absorption system. In comparison to the system shown in Fig. 3, the evaporator (1) and condenser (3) are eliminated. The absorber is replaced by a dehumidifier, which is designed differently to bring into contact the conditioned air with the sorbent solution in order to absorb the humidity directly from the airstream. Similarly, the desorber is replaced by a regenerator, which again brings the sorbent solution to be regenerated into direct contact with ambient or exhaust air. The net result of the operation of this device is the dehumidification of the conditioned airstream, which handles the latent load. To handle the sensible load, the temperature of the conditioned air must be lowered either by an external cooling device or by evaporating water into the now dry airstream using an evaporative cooler in place of the evaporator (1) of Fig. 3.

The basic solar desiccant system as depicted in Fig. 6 can be improved when using liquid desiccants by the addition of a number of heat exchangers for better energy economy. Such a system is shown schematically in Fig. 8. The two main components, the dehumidifier and the regenerator, have a similar design, suitable for contacting the sorbent solution

with air. Both are vertical, packed towers, with the solution sprayed over the packing and flowing down and air brought in at the bottom and flowing up, in counterflow to the solution. A cooling coil inserted in the dehumidifier and employing cooling tower water serves to remove the heat of sorption during dehumidification. Similarly, a heating coil in the regenerator supplies the heat of desorption during regeneration. This is a deviation from the schematic shown in Fig. 6, in which the dehumidifier and regenerator are adiabatic; one of the advantages of a liquid desiccant system is that it allows nonadiabatic dehumidification and regeneration processes, as in closed absorption systems, which makes these processes more efficient. In the system of Fig. 8, outside (fresh) air is dehumidified in the dehumidifier and then sent to the conditioned space. Building exhaust air is preheated in the air-air heat exchanger by regenerator exhaust air and used for solution regeneration. Building exhaust air is preferred to outside air for this purpose since it is dryer. Figure 8 also shows a liquid-liquid heat exchanger between the weak and cool solution leaving the dehumidifier and the strong and hot solution leaving the regenerator for recovering heat from the latter to the former; this is again similar to the recuperative heat exchanger of a closed absorption system. Figure 8 further shows the solar part of the system, consisting of collectors and storage, for supplying heat to the regenerator.

Liquid desiccant systems have many of the same advantages of closed-cycle absorption: They allow

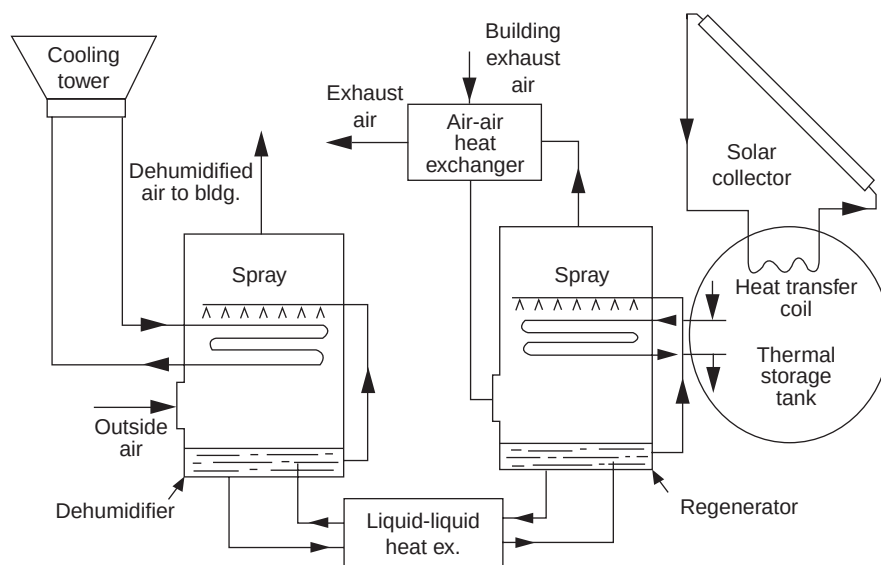


FIGURE 8 Diagram of a liquid desiccant system for dehumidification.

internal heat recovery from the strong and hot solution to the weak and cool solution by means of a recuperative heat exchanger; circulating the sorbent from one unit in the system to the next is easily done by pumping the liquid; and pressure drop in the liquid–air contactors, dehumidifier and regenerator, is relatively low, reducing parasitic losses. The liquid desiccant may be filtered to remove dust and dirt picked up from the air. For solar-powered operation, desorption is possible in a special regenerating collector exposing the desiccant directly to the sun, which can make the regeneration process simpler and more efficient. It also makes it possible to store cooling capacity by storing concentrated liquid desiccant. One disadvantage is the potential carry-over of sorbent in the air, which may lead to sorbent loss and contamination of the conditioned space. There are relatively few liquid desiccant systems available on the market. A limited number of solar-powered air-conditioning systems employing liquid desiccants, mostly experimental, have been constructed during the past three decades.

A promising open-absorption heat pump cycle makes it possible to utilize the solar heat at relatively low temperatures to produce both chilled water and cold and dehumidified air in variable proportions, as required by the load. The cycle, referred to as dehumidifier–evaporator–regenerator (DER), employs environmentally benign liquid desiccants in direct contact with the air. The system typically treats the make-up supply air to handle the latent load and supplies chilled water to handle the sensible load independently. A schematic of the cycle is shown in Fig. 9. It consists of six major components, indicated by roman numerals: an indirect contact evaporative cooler (I), an air dehumidifier or absorber (II), a solution regenerator or desorber (III), two air–air heat exchangers (IV, V), and a solution–solution heat exchanger (VI). Ambient air entering the absorber at state 7 is brought into contact with a strong (concentrated) absorbent solution entering the unit at state 16 and leaving it at state 17. Water vapor is removed from the airstream by being absorbed into the solution stream. The heat of condensation and dilution Q_A released during this absorption process is rejected to a stream of cooling water or ambient air (states 3 and 4). Dehumidified air leaving the absorber at state 8 is drawn through an air–air heat exchanger (IV) into the evaporative cooler. By evaporating externally supplied water into the airstream, the heat Q_E is removed from the chilled medium (states 1 and 2) at a low temperature. The air leaves the evaporative cooler at state 10 humid,

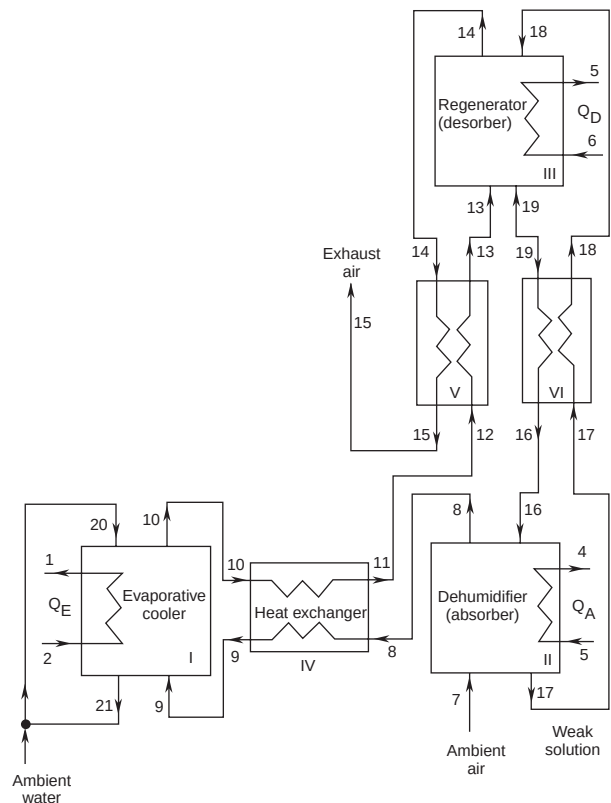


FIGURE 9 Schematic description of the DER open-absorption chiller.

but at a lower temperature than at its inlet, and passes again through the heat exchanger (IV), where it provides precooling for the airstream 8–9 in order to allow for a low cooling temperature. After being preheated in the second air–air heat exchanger (V) by exhaust air, the airstream enters the regenerator at state 13, where it serves to reconcentrate the weak absorbent solution (state 18). The heat of desorption Q_D required to remove water vapor from the solution is provided by an external (solar) heat source (states 5 and 6). The exhaust air leaves the regenerator (state 14), preheats the airstream 12–13 in the air–air heat exchanger (V), and is rejected into the environment. The solution–solution heat exchanger (VI) facilitates preheating of the weak solution (states 17 and 18) and precooling of the strong solution (states 16 and 19).

The advantages of open-cycle DER absorption chillers compared to closed-cycle systems are evident from the previous description. First, the number of main components is reduced by one by transferring condensation of the refrigerant from a condenser to the environment. Second, capital-intensive pressure-sealed units are avoided because the whole system

can be operated at atmospheric pressure. Third, efficient utilization of very low heat source temperatures is possible. The open-cycle DER chiller can be operated with any heat source temperature that is only slightly above the heat rejection temperature. This is particularly favorable for the efficient utilization of heat sources with varying temperatures, such as solar heat from flat-plate collectors. Chilled water at state 21 and/or dehumidified and cooled air at state 9 are the products of the system and may be extracted at varying quantities as required by the air-conditioning application.

5. CONCLUSION

Current trends in solar air-conditioning show a clear preference for absorption cycles over other heat-powered cycles, in two alternative directions: closed-cycle and open-cycle systems. Whereas the main thrust in research on closed-cycle absorption systems has been toward increasing the operating temperature in order to improve efficiency through multi-staging, open-cycle absorption (desiccant) systems have been developed for use with low-temperature heat sources such as flat-plate solar collectors. With both alternatives, the economics of solar cooling are dominated by the solar part of the system—collectors and storage.

Open-cycle systems hold promise for reducing the cost of the solar part of the system and for providing quality air-conditioning with independent temperature and humidity control. Both liquid and solid desiccants are already penetrating the market. Systems such as the DER open-absorption heat pump cycle form the basis for promising cooling systems capable of efficiently exploiting low-tem-

perature solar heat as well as other, similar sources of low-grade heat.

SEE ALSO THE FOLLOWING ARTICLES

Materials for Solar Energy • Refrigeration and Air-Conditioning • Solar Cells • Solar Cookers • Solar Distillation and Drying • Solar Energy, History of • Solar Fuels and Materials • Solar Heat Pumps • Solar Ponds • Solar Thermal Energy, Industrial Heat Applications • Solar Thermal Power Generation • Solar Water Desalination

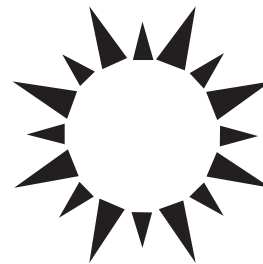
Further Reading

- Alefeld, G. (1982). Regeln für den entwurf von mehrstufigen absorbermaschinen (Rules for the design of multistage absorption machines). *Brennst-Warme-Kraft* 34, 64–73.
- American Society of Heating, Refrigeration and Air-Conditioning Engineers (1998). Absorption cooling, heating and refrigeration equipment. In “ASHRAE Handbook,” refrigeration volume. American Society of Heating, Refrigeration and Air-Conditioning Engineers, Atlanta, GA.
- American Society of Heating, Refrigeration and Air-Conditioning Engineers (2000). Solar energy equipment. In “ASHRAE Handbook,” HVAC systems and equipment volume. American Society of Heating, Refrigeration and Air-Conditioning Engineers, Atlanta, GA.
- Grossman, G. (2002). Solar-powered systems for cooling, dehumidification and air-conditioning. *Solar Energy* 72, 53–62.
- Henning, H. (ed.). (2003). “Solar Assisted Air Conditioning in Buildings—Handbook for Planners.” Springer-Verlag, Vienna.
- Hellmann, H.-M., and Grossman, G. (1995). Simulation and analysis of an open-cycle dehumidifier–evaporator–regenerator (DER) absorption chiller for low-grade heat utilization. *Int. J. Refrigeration* 18, 177–189.
- Herold, K. E., Radermacher, R., and Klein, S. A. (1996). “Absorption Chillers and Heat Pumps.” CRC Press, New York.
- Kreith, F., and Kreider, J. F. (1978). “Principles of Solar Engineering.” McGraw-Hill, New York.

Solar Detoxification and Disinfection

S. MALATO-RODRÍGUEZ

Plataforma Solar de Almería (CIEMAT)
Tabernas, Almería, Spain



1. Solar Photochemistry
2. Photodegradation Principles: Photocatalysis
3. Fundamental Parameters in Photocatalysis
4. Solar Detoxification and Disinfection Technology
5. Solar Detoxification and Disinfection Applications
6. Conclusion

Glossary

biodegradable compounds Compounds that can be metabolized by microorganisms.

compound parabolic concentrator (CPC) Static collectors with a reflective surface following an involute around a cylindrical reactor tube.

conduction band A vacant or only partially occupied set of many closely spaced electronic levels resulting from an array of a large number of atoms forming a system in which the electrons can move freely or nearly so. This term is usually used to describe the properties of semiconductors.

electron/hole Negative and positive charge generated in an illuminated semiconductor.

mineralization Conversion of organic compounds to inorganic ones. Inorganic compounds are usually carbon dioxide, water, and acids (nontoxic substances).

photocatalysis Acceleration of a photoreaction by the presence of a catalyst.

photoreactor Reactor where a photochemical reaction takes place.

semiconductor Material with an appropriate energetic separation between its valence and conduction bands that can be surpassed applying external energy. Metal oxides and sulfides represent a large class of semiconductor materials suitable for photocatalytic purposes.

solar detoxification Photochemical processes driven by solar photons able to destroy toxic substances. The process is based on the generation of hydroxyl radicals in water.

total organic carbon Quantity of organic compounds dissolved in water measured as pure carbon.

The solar heterogeneous photocatalytic detoxification process consists of utilizing the near-ultraviolet (UV) part of the solar spectrum to photoexcite a semiconductor catalyst in the presence of oxygen. In these circumstances, oxidizing species, either bound hydroxyl radical ($\bullet\text{OH}$) or free holes, which attack oxidizable contaminants, are generated, thus producing a progressive breaking of molecules yielding CO_2 , H_2O , and dilute inorganic acids. The most commonly used catalyst is the semiconductor TiO_2 (a cheap, nontoxic, and abundant product). The solar homogeneous photocatalytic detoxification process (photo-Fenton) is based on the production of $\bullet\text{OH}$ radicals by Fenton reagent. This is an extension of the Fenton process, which takes advantage of UV-visible light irradiation at wavelength values higher than 300 nm. During the past decade, thousands of studies and patents related to the heterogeneous photocatalytic removal of toxic and hazardous compounds from water and air have been published, and applications and target compounds are numerous. Where medium to high solar radiation is available, solar detoxification is useful for treating water contaminants with several hundreds of milligrams per liter of nonbiodegradable contaminants and for disinfecting water. This process has proven promising for the treatment of contaminated water, air, and soil. Nevertheless, the specific solar technology employed is not commercially available. This article reviews the use of sunlight to produce $\bullet\text{OH}$ radicals. The experimental systems necessary for performing pilot plant-scale solar photocatalytic experiments are described. The basic components of these pilot plants and the fundamental parameters related to solar photocatalysis reactions are outlined. A description is given of how solar photocatalysis may become a significant segment of the wastewater treatment technologies related to disinfection and the degradation of very persistent toxic compounds.

1. SOLAR PHOTOCHEMISTRY

A specific characteristic of solar photochemical processes is that the photons from the solar spectrum must be absorbed by some component(s) of the reacting system, transferring their energy to the chemical system. Solar photons can be directly absorbed by reactants (direct photochemistry) and/or by a catalyst or sensitizer (photocatalytic processes). Solar detoxification and disinfection are included within photocatalysis.

Photochemical processes use the intrinsic energy of photons (i.e., wavelength) from the solar irradiation to provoke specific reactions. The energy of a photon is a function of its wavelength, according to Eq. (1):

$$E = \frac{hc}{\lambda}, \quad (1)$$

where h is Planck's constant (6.626×10^{-34} Js), c is the speed of light, and λ is the wavelength. Approximately 4 or 5% of the sunlight at the earth's surface is in the near-ultraviolet (UV) region (wavelengths from 300 to 400 nm), approximately 45% is in the visible range (400–760 nm), and the rest is in the near-infrared (IR) and IR regions. Solar photons from the visible and near-UV spectrum are energetic enough to be used in photochemical processes. Infrared photons are normally useless in photochemistry and can be considered as waste heat in chemical processes.

2. PHOTODEGRADATION PRINCIPLES: PHOTOCATALYSIS

A wide range of contaminants to a varying degree resist photolytic, biological, and chemical degradation. In such cases, it is necessary to adopt much more effective systems: air stripping, adsorption on granulated activated carbon, incineration, ozone, and oxidation. Obviously, methods based on chemical destruction, when properly developed, provide complete pollutant abatement, whereas those that only perform phase separation leave a consequent problem of final disposal. Air stripping and adsorption on granulated activated carbon, merely transferring toxic materials from one medium to another, is not a long-term solution to the problem of loading hazardous waste on the environment. Incineration is capable of converting toxic compounds to carbon dioxide, water, and inorganic acids, but negative public perception has slowed and very often prevented its implementation. Concerns over emissions

of undestroyed components and organic products of incomplete combustion have caused incineration to be a source of continuing controversy. Ozone is not very often able to completely destroy pesticides to harmless compounds.

Chemical oxidation aims at the mineralization of the contaminants into carbon dioxide, water, and inorganics. Much research has addressed this goal in the past decade, indicating the prominent role of a special class of oxidation techniques defined as advanced oxidation processes (AOPs), which usually operate at or near ambient temperature and pressure. AOPs, although making use of different reaction systems, are all characterized by the same chemical feature: production of OH radicals ($\bullet\text{OH}$). This radical is strongly oxidative, much more so than other traditional oxidants (e.g., ozone, hydrogen peroxide, chloride dioxide, and chlorine). These can then be used to chemically decompose a pollutant into harmless end products. For example, in aromatic compounds, the aromatic part is hydroxylated and successive steps in oxidation/addition lead to ring opening. The resulting aldehydes and carboxylic acids are decarboxylated and finally produce CO_2 . Methods based on UV, $\text{H}_2\text{O}_2/\text{UV}$, O_3/UV , and $\text{H}_2\text{O}_2/\text{O}_3/\text{UV}$ combinations utilize photolysis of H_2O_2 and ozone to produce the hydroxyl radicals. In addition, other methods, such as heterogeneous photocatalysis or homogeneous photo-Fenton, are based on the use of a wide band gap semiconductor or the addition of H_2O_2 to Fe^{2+} salts and by irradiation with UV-visible light. Both processes are of special interest since sunlight can be used. The main problem with the AOPs is their high cost due to the use of expensive reactants (i.e., H_2O_2) and energy consumption (generation of UV radiation). Therefore, the future applications of these processes may be improved diminishing reactants and energy consumption through the use of catalysis and solar energy.

2.1 Definitions

The term photocatalysis implies the combination of photochemistry with catalysis. Both light and catalyst are necessary to achieve or to accelerate a chemical reaction. Photocatalysis may be defined as the acceleration of a photoreaction by the action of a catalyst. It also refers to a general label to indicate that light and a substance (the catalyst or initiator) are necessary entities for influencing a reaction. Heterogeneous processes employ semiconductor slurries for catalysis, whereas homogeneous photochemistry is used in a single-phase system.

In the case of homogeneous photocatalytic processes, the interaction of a photon-absorbing species (transition metal complexes, organic dyes, or metalloporphyrines), a substrate (e.g., the contaminant), and light can lead to a chemical modification of the substrate. The photon-absorbing species is activated and accelerates the process by interacting through a state of excitation. In the case of heterogeneous photocatalysis, the interaction of a photon produces the appearance of electron/hole pairs, the catalyst being a semiconductor (e.g., TiO_2 and ZnO). In this case, the excited electrons are transferred to the reducible specimen at the same time that the catalyst accepts electrons from the oxidizable specimen that occupies the holes. In both directions, the net flow of electrons is null and the catalyst remains unaltered.

2.2 Heterogeneous Photocatalysis

When a semiconductor is in contact with a liquid electrolyte solution containing a redox couple, charge transfer occurs across the interface to balance the potentials of the two phases. An electric field is formed at the surface of the semiconductor. This means that the two charge carriers should react at the semiconductor/electrolyte interface with the species in solution. The semiconductor-mediated redox processes involve electron transfer across the interface. When electron/hole pairs are generated in a semiconductor particle, the electron moves away from the surface to the bulk of the semiconductor as the hole migrates toward the surface. If these charge carriers are separated fast enough, they can be used for chemical reactions at the surface of the photocatalyst (i.e., for the oxidation or reduction of pollutants).

Metal oxides and sulfides represent a large class of semiconductor materials suitable for photocatalytic purposes. Table I lists selected semiconductor materials, which have been used for photocatalytic reactions, together with band gap energy required to activate the catalyst. The table also indicates the wavelength of radiation required to activate the catalysts.

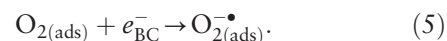
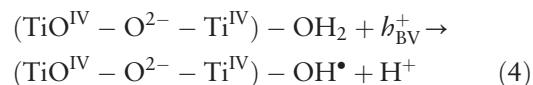
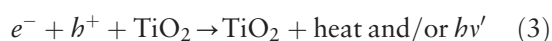
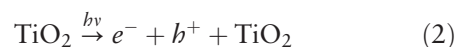
Whenever different semiconductor materials have been tested under comparable conditions for the degradation of the same compounds, TiO_2 has generally been demonstrated to be the most active. TiO_2 's strong resistance to chemical and photocorrosion, its safety, and its low cost limit the choice of convenient alternatives. Furthermore, TiO_2 is of special interest since it can use natural (solar) UV (Table I). Other semiconductors are degraded during the repeated catalytic cycles involved in heteroge-

TABLE I

Selected Properties of Several Semiconductors

Material	Band gap (eV)	Wavelength corresponding to band gap (nm)
SnO_2	3.9	318
ZnS	3.7	336
SrTiO_3	3.4	365
BaTiO_3	3.3	375
ZnO	3.2	380
TiO_2	3.0	390
WO_3	2.8	443
CdS	2.5	497
GaP	2.3	540
Fe_2O_3	2.2	565
CdO	2.1	590
CdSe	1.7	730
GaAs	1.4	887

neous photocatalysis:

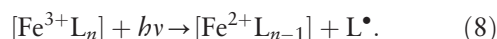
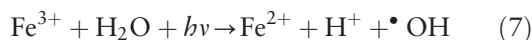
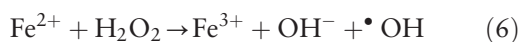


The first effect, after absorption of near-UV radiation ($\lambda < 390 \text{ nm}$), is the generation of electron/hole pairs, which are separated between the conduction and valence bands (Eq. 2). In order to avoid recombination of the pairs generated (Eq. 3), if the solvent is oxidoreductively active (water) it also acts as a donor and acceptor of electrons. Thus, on a hydrated and hydroxylated TiO_2 surface, the holes trap $\bullet\text{OH}$ radicals linked to the surface (Eq. 4). In any case, it should be emphasized that even trapped electrons and holes can rapidly recombine on the surface of a particle (Eq. 3). This can be partially avoided through the capture of the electron by preadsorbed molecular oxygen, forming a superoxide radical (Eq. 5). The oxidative species formed (particularly the hydroxyl radicals) react with the majority of organic substances.

2.3 Homogeneous Photocatalysis

Although the oxidizing power of the combination of Fe^{2+} and H_2O_2 (i.e., the Fenton reaction; Eq. 6) has

been known for more than 100 years, the fact that the reaction can be enhanced by UV-visible irradiation ($\lambda < 580$ nm) was determined only recently. The so-called photo-Fenton reaction (Eq. 7) produces additional hydroxyl radicals and leads to a reduction of the photocatalyst by the irradiated light. Under these conditions, the photolysis of Fe^{3+} complexes allows Fe^{2+} regeneration, and the occurrence of Fenton reactions due to the presence of H_2O_2 and iron may be considered a true catalyst. The main advantage of the photo-Fenton process is the light sensitivity up to a wavelength of 580 nm. The depth of light penetration is high and the contact between pollutant and oxidizing agent is close since a homogeneous solution is used.



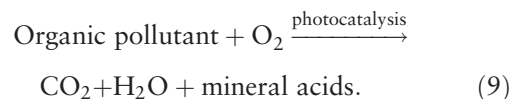
Only a very simplified description is given in Eqs. (6–8). Note that reaction 8, which involves organic ligands L (e.g., RCOO^- , RO^- , and RNH^+), is more efficient than the oxidation of organic substrates via $\bullet\text{OH}$. Therefore, usually contaminants degradation is getting faster during the reaction since the more inert hydrocarbons are first oxidized by $\bullet\text{OH}$ to yield substances containing several hydroxyl or carboxyl groups that are able to participate in reactions thereafter.

3. FUNDAMENTAL PARAMETERS IN PHOTOCATALYSIS

3.1 Influence of Oxygen and pH

Whatever the formation pathway, it is well-known that O_2 and water are essential for detoxification. There is no degradation in the absence of either. The pollutants are usually organic; therefore, Eq. (9) summarizes the overall process. Given the reaction stoichiometry of this equation, there is no photo-mineralization unless O_2 is present. Oxygen is necessary for complete mineralization and does not seem to be competitive with other reactives during the adsorption on TiO_2 since the places where oxidation takes place are different from those of reduction. The concentration of oxygen also affects the reaction rate, but the difference between using air ($p\text{O}_2 = 0.21$ atm) or pure oxygen ($p\text{O}_2 = 1$ atm) is not drastic. Oxygen also intervenes, governing the efficiency of a photocatalytic oxidative degradation,

thus minimizing electron/hole recombination by capturing the photogenerated electron (see Eq. 5).



The oxide/electrolyte interface has an electrical surface charge that strongly depends on the pH of the medium. The pH of the aqueous solution significantly affects TiO_2 , including the charge of the particle and the size of the aggregates it forms. The zero surface charge (PZC) yields zero electrostatic surface potentials that cannot produce the interactive rejection necessary to separate the particles within the liquid. This induces a phenomenon of aggregation and TiO_2 clusters become larger. This effect is clearly related to the capability of the suspension to transmit and/or absorb light. Furthermore, larger clusters sediment more quickly than small particles; thus, the agitation necessary to maintain perfect homogeneity must be more vigorous. The PZC for TiO_2 is approximately 7. Above and below this value, the catalyst is negatively or positively charged according to $\text{TiOH} \geq 80\%$ when $3 < \text{pH} < 10$, $\text{TiO}^- \geq 20\%$ when $\text{pH} > 10$, and $\text{TiOH}_2^+ \geq 20\%$ when $\text{pH} < 3$. Under these conditions, the photocatalytic degradation of ionizable organic compounds is affected by the pH. At first sight, a very acidic solution appears to be detrimental and a very basic solution to be favorable since the variations are modest or nonexistent around neutrality. In many cases, a very important feature of photocatalysis is not taken into account when it is to be used for decontamination of water: That is, during the reaction, a multitude of intermediate products are produced that may behave differently depending on the pH of the solution. To use only the rate of decomposition of the original substrate could yield an erroneous pH as the best for contaminant degradation. Therefore, a detailed analysis of the best pH conditions should include not only the initial substrate but also the rest of the compounds produced during the process.

3.2 Catalyst Concentration Influence

Incident radiation on the reactor and length of path of light inside the reactor are fundamental in determining optimum catalyst concentration. If the radiation comes from a source outside the photo-reactor (as in a reactor illuminated by solar radiation) and the path length is short (1–2 cm maximum), maximum reaction rate is obtained with 1 or 2 g/liter

of TiO_2 . If the path length is several centimeters long, the appropriate catalyst concentration is several hundred milligrams per liter. The major difference between considering the TiO_2 from dry powder and TiO_2 particles suspended in an aqueous medium involves the mean size of the particles. If two samples with the same catalyst concentration but that are dispersed with different protocols (ultrasounds, stirring, etc.) are compared; sonicated samples are found to be more efficient. Although in stirred samples there are fewer TiO_2 clusters, larger clusters screen light better than small ones. It has been demonstrated that efficiencies are better with small particles than with larger particles only when reactor diameter and catalyst concentration are optimized to allow UV photons to penetrate along the entire photoreactor path length. In this case, it is not possible to talk about particle radius because TiO_2 powder is irregularly shaped, and when it is dispersed in an aqueous medium the “particles” (300–600 nm) are clusters of primary particles (20–40 nm), not spherical or monodisperse. This leads to the conclusion that light extinction in colloidal suspensions is a determining parameter for solar photoreactor design. It should also be recalled that small particle sizes cause additional problems for catalyst separation after photocatalytic treatment. The best inner reactor diameter for solar photocatalytic applications is in the range of a few centimeters with a few hundred milligrams of TiO_2 per liter. The area:volume ratio is also a crucial parameter, and if this ratio is optimized the reactor efficiency would also be increased.

Because iron acts as a catalyst in the photo-Fenton reaction, the degradation rate should increase with the increasing amount of iron added, but it reaches a maximum at the optimum iron concentration. Further addition of iron does not increase the degradation rate. This threshold is dependent on the light source, the thickness of the irradiated layer, and the type of wastewater because internal light filtering may occur. For solar reactors, it is therefore necessary to determine the optimum iron concentration experimentally—that is, the minimum concentration at which the maximum reaction rate is obtained. When catalyst concentration is very high, after traveling a certain distance on an optical path length, turbidity impedes further penetration of light in the reactor. The percentage of photons absorbed by the suspension and the percentage of photons scattered is a very complex problem that cannot be solved experimentally but must be experimentally estimated.

3.3 Influence of Initial Contaminant Concentration

The degradation rate observed for an organic substrate follows saturation behavior. After a certain concentration is achieved, the rate increases very little and in some cases a decrease is observed. If possible, the optimum contaminant concentration in water before the photocatalytic treatment should permit the maximum reaction rate. Due to the kinetics of the photocatalytic reaction, the rate decreases considerably when contaminant concentration is low. The photocatalytic disappearance of contaminants with TiO_2 follows apparent first-order kinetics when the initial concentration is low enough and no catalyst saturation occurs. It should be emphasized that contaminants decompose, giving rise to intermediates, which could also be competitive on the surface of the TiO_2 . Their concentration varies throughout the reaction until their mineralization, and thus Eq. (10) (based on the Langmuir-Hiselswood model) may describe the kinetics:

$$r = \frac{dC}{dt} = \frac{krKC}{1 + KC + \sum_{i=1}^n K_i C_i (i = 1, n)}, \quad (10)$$

where k_r is the reaction rate constant, K is the reactant adsorption constant, C is the concentration at any time, K_i is the adsorption constant of the intermediates, and C_i is their concentration at any time. When C_0 is low enough, Eq. (10) can be simplified ($1 + KC + \Sigma \dots = 1$) to an order one-reaction rate equation ($r = k_{ap}C$). Likewise, at higher concentrations Eq. (10) can be simplified by adjusting it to a zero-order one: $(C_0 - C) = k_r t$.

During the treatment of real wastewaters, since the reaction is not expected to follow simple models such as first- or zero-order kinetics, very often overall reaction rate constants cannot be calculated. Total organic carbon (TOC) analysis is highly recommended for this because of its versatility and, at the same time, its suitability for “on-line” treatment plant installations. Mineralization (TOC disappearance) only occurs when the last step of the oxidation process yields CO_2 . The maximum gradient of the mineralization curve, $\text{TOC} = f(t)$, which is the gradient of the tangent at the inflection point ($r_{\text{TOC}, 0}$), is employed to obtain a practical point of comparison for different experiments since it has a zero-order rate constant (mg/liter/min) and therefore appears easy to handle. This parameter ($r_{\text{TOC}, 0}$) is referred to as the maximum rate.

Once the optimum initial concentration for degradation is known, a model for predicting plant

behavior is necessary. This model should enable one to calculate the solar collector area required for treating water contaminated with different amounts of pesticides. A developed model, which has the analytic form of a Langmuir–Hinshelwood equation, has been preferred for fitting experimental data in large solar photocatalytic plants by approximate solution of the general photocatalytic kinetic system. With these considerations, the rate of TOC disappearance is given by

$$r_{\text{TOC}} = \frac{\beta_1 [\text{TOC}]}{\beta_2 + \beta_3 [\text{TOC}]} \quad (11)$$

Experimental results can be used to calculate the constants (β_i). Equation (12) enables one to predict TOC degradation as a function of the initial TOC and of the available time (or radiation energy) and, reciprocally, to predict the incident energy on the reactor necessary to reach a specific degree of mineralization. Nevertheless, UV radiation data for the final plant location must be available.

$$\frac{1}{\beta_1} \left\{ \beta_2 \ln \left(\frac{[\text{TOC}]_0}{[\text{TOC}]} \right) + \beta_3 ([\text{TOC}]_0 - [\text{TOC}]) \right\} = t. \quad (12)$$

3.4 Influence of Radiant Flux

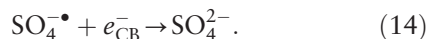
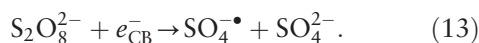
Since 1990, there has been clarification of the kind of solar technology that should be involved in detoxification and disinfection. The question was whether concentrating the radiation is really necessary for the photocatalysis technology and whether a nonconcentrating collector can be as efficient as concentrating ones. Initially, it was thought that concentrating collectors were the ideal alternative and, in fact, the first large pilot plants use them. However, their high cost and the fact that they can only operate with direct solar radiation (implying their location in highly insolated areas) led to consideration of static nonconcentrating collectors. Early engineering-scale tests with solar systems were based on parabolic troughs with concentrations in the range of $10\times$, but a second generation of nonconcentrating collectors soon began to be tested, including thin films, flat plates, and compound parabolic collectors (CPCs). Use of one-sun systems for water treatment is based on two factors: the high percentage of UV photons in the diffuse component of solar radiation and the dependence of rates on light intensity.

It was shown experimentally that above a certain flux of UV photons, the reaction rate changes from

one to a half order depending on the intensity. This modification does not seem to happen at a specific radiation intensity because different researchers have obtained different results. Presumably, the experimental conditions have a significant effect. Most authors attribute the transition of $r=f(I^{1.0})$ to $r=f(I^{0.5})$ to excess photogenerated species (e^- , h^+ , and $\bullet\text{OH}$). Because the reaction rate depends on the amount of hydroxyl radicals present, and these are generated in the holes (see Eq. 4), $r \propto I^{0.5}$ when I is high. At higher radiation intensities, another transition from $r=f(I^{0.5})$ to $r=f(I^0)$ occurs. At this moment, the photocatalytic reaction ceases to depend on the received radiation and instead depends only on the mass transfer within the reaction. Thus, the rate is constant even though the radiation increases. This effect can be attributed to different causes, such as the lack of electron scavengers (i.e., O_2) and organic molecules in the proximity of the TiO_2 surface and/or excess products occupying active centers of the catalyst. These phenomena occur more frequently when working with supported catalyst and/or at a low agitation level. This implies a low catalyst surface in contact with the liquid and smaller turbulence. This does not favor the contact of reactants with the catalyst and the diffusion of products from the proximity of the catalyst to the liquid.

These effects may be appreciably attenuated if a product that reduces the importance of the electron/hole recombination is added. When the electrons are trapped, recombination of e^-/h^+ is impeded. Either way, the addition of oxidants can improve the efficiency of the process at high illumination intensities. Moreover, oxidants can increase the quantum yield even at low irradiation levels due to their strong oxidizing character. The use of inorganic peroxides has been demonstrated to enhance the rate of degradation of different organic contaminants because they trap the photogenerated electrons more efficiently than O_2 . The increased photocatalytic reaction rate with these additives would decrease photoreactor dimensions proportionally and dramatically decrease overall costs. The oxidizing substance should not generate any toxic by-product, so hydrogen peroxide is the obvious candidate. The effect depends on H_2O_2 concentration, generally showing an optimum range of concentration. Peroxydisulfate can also be a beneficial oxidizing agent in detoxification because it reacts with the semiconductor photogenerated electrons (e_{CB}^-). The peroxydisulfate ion accepts an electron and dissociates (Eq. 13). The sulfate radical proceeds through the reaction shown in Eq. (14). In addition, the strongly oxidizing $\text{SO}_4^{\bullet-}$ ($E^\circ = 2.6 \text{ V}$) can

directly participate in the degradation processes.



4. SOLAR DETOXIFICATION AND DISINFECTION TECHNOLOGY

4.1 Collectors for Solar Water Detoxification

CPCs are a good option for solar photochemical applications. CPCs are static collectors with an involute reflective surface around a cylindrical reactor tube. They have been found to provide the best optics for systems with a low solar radiation concentration and can be designed with a concentrating ratio close to 1 for capturing both direct and diffuse UV sunlight. The reflector design enables almost all the UV radiation arriving at the CPC aperture to be collected and available for processing in the reactor. The light reflected by the CPC is distributed around the back of the tubular photo-reactor so that most of the reactor tube circumference is illuminated. It is easy to pipe water in and to distribute it. Most of the components of solar photochemical systems are made of standard materials with no special requirements except for the reactor, the reflective surface, and the catalyst.

4.2 Reflective Surfaces

The optical quality requirements of reflective surfaces for solar applications are usually related to the concentration required by the particular application under consideration. The higher the concentration desired, the stricter the requirements for quality of parameters. In the case of solar photochemical applications, the strictest requirements are for parabolic trough collectors (PTC); for example, UV-mirror materials must have a specular reflectance between 300 and 400 nm to achieve concentration ratios from 1 to 20. The larger the errors, particularly the reflective surface errors, the lower the effective concentration ratio. The reverse is also true: The lower the effective concentration ratio, the higher the optical errors may be, and therefore the lower the quality of reflective surface required. This is an important characteristic favoring low or nonconcentrating systems since these lower quality requirements (lower specular reflectance) directly

translate into lower manufacturing cost; the reflector element can represent a considerable fraction of collector cost. With regard to the reflector/concentrator, another important factor is the reflective base material. The majority of solar photochemical processes require a highly reflective material for UV radiation. Reflectivity (reflected radiation/incident radiation) between 300 and 400 nm for traditional silver-coated mirrors is very low, so aluminum-coated mirrors are the best option. The ideal reflective surface for solar photochemical applications must have high reflectivity in the UV range, acceptable durability under outdoors conditions for extended service lifetimes, and a reasonable cost. The surfaces that best fit these requirements are electro-polished-anodized aluminum (electrolytically formed aluminum oxide outer layer) and organic plastic films with an aluminum coating.

4.3 Photocatalytic Reactor

The photochemical reactor must contain the working fluid, including the catalyst or sensitizer, and must transmit solar UV light efficiently with minimal pressure drop across the system. When a large array of solar collectors are going to be used, such as in the case of treatment of water contaminants by solar photocatalysis, the reactor must be hard enough to work under usable water pressure. Tube configurations are clearly most appropriate for fluid containment and pumping when large volumes are to be processed. The choice of materials that are both transmissive to UV light and resistant to its destructive effects is limited. Photochemical reactors must be able to withstand summer temperatures of approximately 70–80°C. In addition, reactor material must be inert with regard to the chemicals that must be contained. Common materials that meet these requirements are fluoropolymers, acrylic polymers, and several types of glass. Quartz has excellent UV transmission, but the advantage in transmission in the terrestrial solar spectrum compared to other materials does not justify its high cost. Fluoropolymers are a good choice for photo-reactors due to their good UV transmittance, but one of their major disadvantages is that in order to achieve a desired minimum pressure rating, the wall thickness of a fluoropolymer tube may have to be increased, which in turn will lower its UV transmittance. Glass is a natural alternative for photoreactors. Standard glass, used as a protective surface, is not satisfactory because it absorbs part of the UV radiation that reaches it due to its iron content. Borosilicate glass has good transmissive properties in the solar range with a cutoff

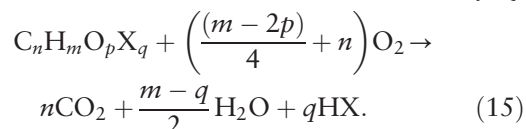
of approximately 285 nm. Therefore, such a low-iron content glass would seem to be the most adequate. Because both fluoropolymers and glass are valid photoreactor materials, cost is an important issue.

5. SOLAR DETOXIFICATION AND DISINFECTION APPLICATIONS

5.1 Wastewater Treatment

The hydroxyl radical ($\cdot\text{OH}$) is the major intermediate reactive responsible for organic substrate oxidation. The free radical $\text{HO}_2\cdot$ and its conjugate $\text{O}_2^{\cdot-}$ are also involved in degradation processes but are much less reactive than free hydroxyl radicals. These radicals react strongly with most organic substances by hydrogen abstraction or electrophilic addition to double bonds. Free radicals further react with molecular oxygen to give a peroxy radical, initiating a sequence of oxidative degradation reactions that may lead to complete mineralization of the contaminant. In addition, hydroxyl radicals may attack aromatic rings at positions occupied by a halogen, generating a phenol homologue. The electrons of the conduction band can also degrade organic compounds by reductive pathways.

In general, the types of compounds that can be degraded include alkanes, haloalkanes, aliphatic alcohols, carboxylic acids, alkenes, aromatics, haloaromatics, polymers, surfactants, pesticides, and dyes. Equation (15) generally holds true for a halogenated organic compound of the general formula $\text{C}_n\text{H}_m\text{O}_p\text{X}_q$:



In photodegradation, transformation of the parent organic compound is desirable in order to eliminate its toxicity and persistence, but the principal objective is to mineralize all pollutants. Furthermore, the stoichiometry proposed for the general reaction (Eq. 15) has to be demonstrated in each case by a correct mass balance. Reactives and products might be lost, producing unreliable results. Mineralization rate is determined by monitoring inorganic compounds, such as CO_2 , Cl^- , SO_4^{2-} , NO_3^- , and PO_4^{3-} . When organics decompose, a stoichiometric increase in the concentration of inorganic anions is produced in the treated water, and very often there is an increase in the concentration of hydrogen ions (decrease in pH). In order to demonstrate that there are no product losses, the molar ratio must be in accordance with the organic

substrate structure. The absence of total mineralization has been observed only in the case of *s*-triazine herbicides, for which the final product obtained is essentially 1,2,5-triazine-2,4,6-trihydroxy (cyanuric acid), which is not toxic. This is due to the strong stability of the triazine nucleus, which resists most methods of oxidation. Nitrogen-containing molecules are mineralized into NH_4^+ and NO_3^- . Ammonium ions are relatively stable and the proportion depends mainly on the initial oxidation degree of nitrogen and on the irradiation time. Analyses of aliphatic fragments resulting from the degradation of the aromatic ring have mainly revealed formate and acetate ions. Other aliphatics (presumably acids, diacids, and hydroxylated compounds) have also been found. Formate and acetate ions are rather stable, which in part explains why total mineralization takes much longer than dearomatization.

Solar photocatalytic mineralization of organic pollutants in water, using the interaction of UV radiation and semiconductor catalysts, has major potential in the industrial destruction of toxic organics in water, and the number of applications and target compounds are numerous. In areas with medium to high solar radiation, solar detoxification is useful for treating water with a maximum organic concentration of several hundred milligrams per liter, nonbiodegradable contaminants, and hazardous contaminants present within complex mixtures of organics. As a result of intense research carried out by groups all over the world during the past 20 years, solar detoxification is a good solution with many interesting applications. During the past decade, there have been thousands of publications and patents on heterogeneous photocatalytic removal of toxic and hazardous compounds from water and air. Some examples are phenols and chlorophenols, chlorinated hydrocarbons, pharmaceutical compounds (antibiotics, antineoplastics, and other hazardous pharmaceutical waste), agrochemical waste (pesticides), and cyanides. They are produced in industrial quantities in processes such as the manufacture of phenol-formaldehyde resins, volatile organic compound (VOC) scrubbers, PVC production, wood-preserving waste, cleaning of marine tank terminals and pipelines, and manufacture of pesticides and rinse water from the equipment used for their application.

Here, pesticides are used as an example for demonstrating the practical application of solar photocatalysis. Although pesticides, and agrochemical compounds in general, have been detected in water since the 1950s and 1960s, in the past 15 years their use has increased dramatically throughout the

world and has nearly doubled every 5 years since 1975. The United Nations estimates that of all pesticides used in agriculture, less than 1% actually reach the crops. The remainder ends up contaminating land, air, and especially water. Therefore, pesticide pollution of water is one of today's greatest environmental problems, with widespread ecological consequences. In areas where agriculture is intensive, the following are major sources of water pollution:

Pesticide treatment as a routine agricultural practice: Approximately 0.2 tons/ha are usually applied in intensive agriculture by classical spraying or in the irrigation pipeline.

Rinse water polluted with pesticides from containers and spray equipment: Small amounts of pesticide remaining in empty containers (approximately 70 units/ha) are typically spilled out into the environment.

Wastewater from the agricultural industry: Water from cleaning or postharvest treatment of fruits and vegetables before packing produces approximately 1 m³/day/ha of wastewater.

Plant waste contaminated with pesticides: It is estimated that 30 tons/ha/year of untreated plant waste is normally disposed of in landfills.

One of the major problems is the uncontrolled dumping of empty pesticide containers. Although the amount of pesticide remaining in each container is very small, millions of them are dumped every year, making it one of the most important sources of pollution in areas with this type of agriculture. A proposed solution is their selective collection and transportation to a recycling plant, where they are rinsed for later reuse. The resulting rinse water, contaminated by the different mixtures of pesticides, must be treated. Simple, inexpensive, and accessible technologies for the *in situ* treatment of such water are therefore necessary. Herbicides are traditionally removed from water using granular or powdered-activated carbon, nanofiltration, ozonation, etc., but these processes have inherent limitations in applicability, effectiveness, and costs. On the other hand, the problems associated with their disposal are quite well suited to solar photocatalytic treatment for the following reasons: (i) The initial pesticide concentration can be controlled as a function of the number of containers washed with the same water, so the most appropriate concentration for optimum photocatalytic efficiency can be chosen; (ii) toxicity is extreme, low volume, and in a well-defined location; (iii) such point sources of pollution may be ideally treated in small-scale treatment units; and (iv) intensive agriculture in greenhouses is usually concentrated in sunny countries.

Solar detoxification has also demonstrated efficiency in the degradation of the halogenated solvents found in much of pharmaceutical industry waste and VOC emissions. Environmental regulations and directives are pushing industrial managers to control VOC emissions. One of the VOC emission control methods is wet scrubbing, but the contaminated water from the scrubbers has to be treated. This could easily be done by solar detoxification. Another source of halogenated wastes is halocompound fabrication. The waste can be estimated as a low percentage of total production, dissolved at 100–200 mg/liter. PVC, for example, produces 2.5 m³ of wastewater effluent contaminated with short-chain polymers or PVC monomer for each ton of PVC. A feasibility study determined that a 6-m² photoreactor can completely mineralize 100 mg/liter of chloroform in 250 liters of wastewater in 8 h of sunlight.

Decontamination of water effluents is one of the most successful photochemical applications of solar photons. The technology has been validated with the construction of large pilot plants. Solar photocatalysis promises to be an important breakthrough in the implementation of solar processes as a specific solar technology for water treatment becomes commercially available. Market analyses show that this environmentally friendly technology, which can destroy many of the most problematic persistent organic pollutants, has numerous potential applications. There can be no doubt that photocatalytic water detoxification is full of promise, as demonstrated by the progress made during the past 10 years.

However, an analysis of the historical evolution of solar photocatalysis clearly identifies three different stages of development. Initial efforts of solar-conscious researchers focused on transferring laboratory research to solar-engineered testing with existing technology. These first results produced excitement in the photochemical research community. Their extrapolation to practical situations assumed an ability to degrade almost any organic contaminant as well as some metals. Later, more appropriate collectors and designs were developed, but the need to know the fundamentals of certain aspects of the reaction led to an increasing number of studies on kinetics, mechanisms, performance of mixtures, and operating parameters with mixed results. It was a period of both promising and discouraging results. In the third stage, which seems to be under way, the boundary conditions of applications are being determined, and the technology is focusing on a specific initial applications, with the particularity that early development and unsolved questions

coexist with near-commercial and industrial applications of the technology. As a result, the environmental market, although very receptive to clean energy resources, is reluctant to accept such a "risk-assuming" initiative.

5.2 Water Disinfection

In many areas of developing countries, cross-contamination of the water supply with both chemical and microbiological contaminants is a major problem; water decontamination and disinfection is recognized by the World Health Organization as one of the most important barriers for human health protection. Solving the problem of trace levels of hazardous nonbiodegradable contaminants is very difficult in developing countries. However, microbial contamination of drinking water due to the lack of treatment sewage or to inadequate working conditions remains the principal water-quality problem in these countries. Access to clean water is restricted by the contamination of water resources, the lack of adequate storage facilities, and the absence of effective delivery systems. Chlorine is widely used as a main disinfectant, but reports show that overall water service coverage in rural areas is low because of many factors (technological, cost, maintenance, social, cultural, logistic, education, etc.). In addition, there is risk caused by the chlorinated by-products, such as trihalomethane and other total organohalides. Chlorine, the most common treatment, is best suited for broad municipal water supply projects for which it is relatively easy to control and monitor the dosage level of disinfectant. However, in the smaller rural and marginalized communities, chlorine use can be impractical or even dangerous. Without proper dosage regulation, chlorine overdoses can cause carcinogenic by-products and strong odors and tastes. Other existing technologies, such as ozone and UV disinfection, are clearly very difficult to implement in rural areas. UV light and ozonation are particularly unsuitable for isolated areas because of the amount of complicated equipment necessary and, in the case of ozone, the danger of toxic waste leakage.

The use of solar energy as a primary alternative to chlorination could prove to be economically viable in countries with a high degree of sunlight. Solar treatment could also be an efficient way to degrade organic matter, making chlorinated drinking water harmless for human consumption. Photocatalytic oxidation is a promising technology for the disinfection of water and wastewater because the $\cdot\text{OH}$ radical is highly toxic toward microorganisms.

6. CONCLUSION

Although its development presents difficulties, treatment of industrial wastewater seems to be one of the most promising fields of application of solar detoxification. The only general rule is that there is no general rule; each case is completely different from any other. Consequently, preliminary research is almost always required on a case-by-case basis to assess and optimize potential pollutant treatment for any problem.

SEE ALSO THE FOLLOWING ARTICLES

Solar Distillation and Drying • Solar Fuels and Materials • Solar Heat Pumps • Solar Thermal Energy, Industrial Heat Applications • Solar Thermal Power Generation • Solar Water Desalination • Sun, Energy from

Further Reading

- Ajona, J. A., and Vidal, A. (2000). The use of CPC collectors for detoxification of contaminated water: Design, construction and preliminary results. *Solar Energy* 68(1), 109–120.
- Blake, D. M. (1999). Application of the photocatalytic chemistry of titanium dioxide to disinfection and the killing of cancer cells. *Sep. Purif. Methods* 28(1), 1–50.
- Cassano, A. E. (2000). Reaction engineering of suspended solids heterogeneous photocatalytic reactors. *Catalysis Today* 58, 167–197.
- Chiron, S., Fernández-Alba, A., Rodríguez, A., and García-Calvo, E. (2000). Pesticide chemical oxidation: State of the Art. *Water Research* 34(2), 366–377.
- Helz, G. R., Zepp, R. G., and Crosby, D. G. (eds.). (1994). "Aquatic and Surface Photochemistry." Lewis, Boca Raton, FL.
- Hoffmann, M. R., Martin, S. T., Choi, W., and Bahnemann, D. (1995). Environmental applications of semiconductor photocatalysis. *Chem. Rev.* 95, 69–96.
- Legrini, O., Oliveros, E., and Braun, A. M. (1993). Photochemical processes for water treatment. *Chem. Rev.* 93, 671–698.
- Malato, S., Blanco, J., and Herrmann, J. M. (eds.). (1999). Solar catalysis for water decontamination. *Catalysis Today* 54(2–3).
- Malato, S., Blanco, J., Vidal, A., and Richter, C. (2002). Photocatalysis with solar energy at a pilot-plant scale: An Overview. *Appl. Catal. B Environ.* 37, 1–15.
- Safarzadeh-Amiri, A., Bolton, J. R., and Cater, S. R. (1996). The use of iron in advanced oxidation processes. *J. Adv. Oxid. Technol.* 1(1), 18–26.
- Serpone, N., and Emeline, A. V. (2002). Suggested terms and definitions in photocatalysis and radiocatalysis. *Int. J. Photoenergy* 4, 91–131.
- Serpone, N., and Pelizzetti, E. (eds.). (1989). "Photocatalysis: Fundamental and Applications." Wiley, New York.

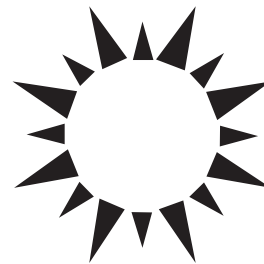
Solar Distillation and Drying

SALEH AL-KHARABSHEH

Hashemite University
Zarqa, Jordan

D. YOGI GOSWAMI

University of Florida
Gainesville, Florida, United States



-
1. Solar Distillation Background
 2. Solar Distillation Processes
 3. Solar Still
 4. Still Performance
 5. Operation and Maintenance
 6. Solar Drying: Definition and Background
 7. Drying Processes
 8. Solar Drying Processes
 9. Solar Dryers
 10. Dryer Performance

Glossary

absorptivity A property of a surface, defined as the fraction of incident energy absorbed on the surface.

convection A mode of heat transfer in which heat is transferred due to the motion of the fluid.

dessicant A material that absorbs moisture from air.

dew point The temperature at which water vapor in air starts to condense when cooled.

drying capacity The quantity of products that can be dried in a single batch of loading.

drying potential The ability of air to evaporate water from the material to be dried, which depends on the temperature and relative humidity of the air.

drying ratio The ratio of the weight of wet products entering the dryer to the weight of the products leaving the dryer.

humidity ratio The ratio of the mass of water vapor in the air to the mass of the dry air.

liner Any natural black or painted material that can withstand the operating temperature and saline water.

natural convection The motion of the fluid due to buoyancy forces, which arise from density differences, caused by temperature variations in the fluid.

regeneration The process of removal of water from the desiccant.

relative humidity The ratio of the mass of water vapor in a given volume of air to the maximum amount of vapor that the same volume of air can hold at the same temperature and pressure.

saturation The condition at which a fluid holding another substance cannot hold any additional amount at the given conditions.

scale The deposition of minerals on any solid surface.

solubility limit The maximum amount of solid that can be dissolved in a liquid at a specified temperature.

sublimation The phase change of a material directly from solid to gas without passing through a liquid phase.

vapor pressure The actual pressure exerted by vapor.

Solar distillation is the process in which solar heat is used to purify water from an impure water source by evaporation and condensation. When solar distillation is used to purify water from saline water, the process is also called solar desalination. Desalination converts saline water with high salt content, about 3.5% by weight in seawater and about 0.6% in brackish water, into fresh water suitable for drinking and other purposes. Solar energy can be used for desalination, either as thermal energy through the use of solar thermal collectors or solar ponds, or as electricity, mainly through the use of photovoltaic cells.

1. SOLAR DISTILLATION BACKGROUND

Solar distillation has been practiced for a long time. The first large-scale solar distillation plant was constructed in 1872 at Las Salinas, Chile, and

consisted of solar stills of about 4700 m² total area, producing about 23,000 l/day. Solar distillation gained more attention after World War I, when many devices were developed. Solar distillation is suitable for remote, arid, and semiarid areas, where drinking water shortage is a major problem and solar radiation is high. The drawbacks to using solar energy for distillation are the high initial cost and the intermittent nature of the sun. Due to these limitations, the present capacity of solar desalination systems worldwide is about 0.01% of the existing large-scale conventional distillation plants. Those plants are small- to medium-capacity installations serving a small number of people and poor communities.

2. SOLAR DISTILLATION PROCESSES

Solar energy can be used to supply the required energy for a desalination process either in the form of thermal energy needed for distillation or as electricity that can be used to drive processes requiring electricity. Solar desalination processes can be classified into two categories:

1. *Indirect systems.* These are usually of small or medium capacity and involve two separate subsystems, a solar energy system and a conventional plant for using the collected solar energy to produce fresh water.

2. *Direct systems.* Here the heat collection and distillation processes take place in the same equipment. In such systems, solar energy is used to produce the distillate directly in the solar still. The still acts as a trap for solar radiation (the greenhouse effect). As solar radiation passes through the transparent cover, it gets absorbed by the water and the absorber surface. As a result the water temperature, thus the vapor pressure, increases and becomes greater than that of the cover and the water evaporates. The vapor rises to the cover by natural convection where it is condensed on the inner side of the cover. A gravity feed funnels the condensate into the collection tank.

3. SOLAR STILL

The most common solar still is a greenhouse type. Greenhouse solar stills are simple to construct and operate with a transparent cover, which may be symmetrical (Fig. 1) or asymmetrical. The main parts of a greenhouse solar still are described next:

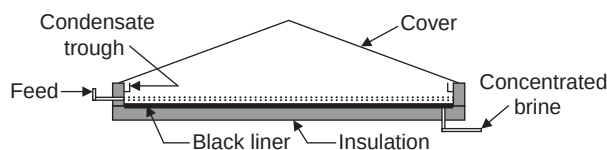


FIGURE 1 Schematic of a symmetrical greenhouse solar still.

Basin. Saline water is kept for distillation in the basin. The basin may be insulated from the bottom to minimize heat loss to the ambient. It is painted black or lined with a black sheet to maximize its absorptivity. The basin may range from 10 to 20 mm deep (shallow basin) to 100 mm or more (deep basin). The basin floor must be inclined slightly, around 1°, to allow for easy drainage of the concentrated brine. The materials needed to construct a still must be resistant to saline water corrosion, and preferably be inexpensive, durable, and available locally. Aluminum, concrete, wood impregnated with a resin, iron painted to resist corrosion, or plastic may be used.

Transparent cover. The cover acts as the condenser, so that the vapor condenses on its inside surface and collects at the bottom. It also reduces heat losses from the hot water by keeping the wind away from it. The slope of the cover must be steep enough to allow the water to flow all the way down to the collection troughs, without falling back into the basin. The cover is usually made of glass or plastic, though glass is preferred because of its high transparency, rigidity, and comparatively low cost. One disadvantage of glass is that it breaks on impact. Plastics, on the other hand, are lightweight, inexpensive, less likely to break, and easy to handle. Their main disadvantages are discoloration and deterioration due to ultraviolet (UV) radiation in the sunlight.

Collection trough. The collection troughs are placed at the lower edges of the cover to collect the distillate. Troughs are pitched along its length so that the distillate flows to the lower end of the still, from where it is collected.

The design of a solar still requires optimization of many factors: brine depth, thermal insulation, cover slope, shape, and material for the still. A well-designed still must also have adequate provision to collect the rainwater that falls on it, so that the surrounding ground is not eroded or flooded.

A variety of basin type solar still designs are used in practice. They differ from each other in the type of

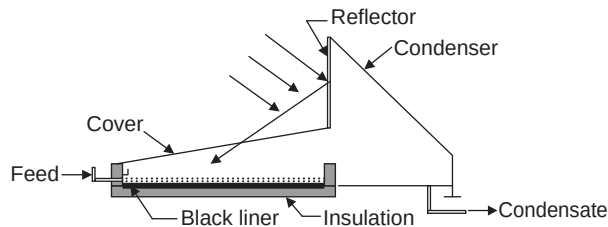


FIGURE 2 Schematic of a greenhouse solar still with a reflector.

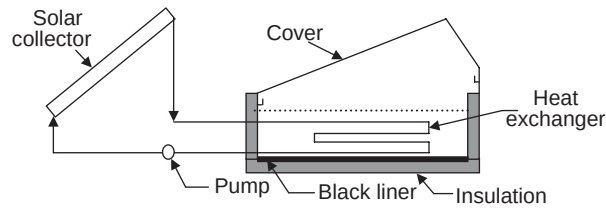


FIGURE 3 Schematic of a greenhouse solar still with a solar collector.

materials used, geometry, the method of supporting the transparent cover, and the arrangements for supply and discharge water. The incident solar radiation on the still can be increased by using a reflector (Fig. 2), which focuses it, so that the still receives both the reflected as well as incident radiation. The temperature of the saline water inside the still may be increased by connecting the still to a solar collector (Fig. 3), or the saline water may be preheated in the collector before entering the still. The latent heat of condensation may be partially utilized by means of a multi-effect solar still (Fig. 4). As the vapor from one stage condenses at the lower surface of the next stage, it gives up its latent heat at that stage thus evaporating a part of the water there. The condensate flows toward the trough and is collected as the output. Thin film evaporation creates a higher output than pool evaporation. This idea may be used in a tilted solar still, as shown in Fig. 5. The still is tilted to face the sun and the feed water flows from the top inlet toward the bottom. The flow rate needs to be properly controlled to generate the optimum output. Other designs include a tilted wick, which looks like a tilted solar still, an inflated solar still where a plastic sheet is used as the transparent cover, and a still with an internal condenser to enhance the condensation process and to preheat the feed water. Stills based on the humidification-dehumidification principle have also been developed. In a humidification-dehumidification still, saline water is heated in a collection system and pumped to the top of a humidification column. Air is

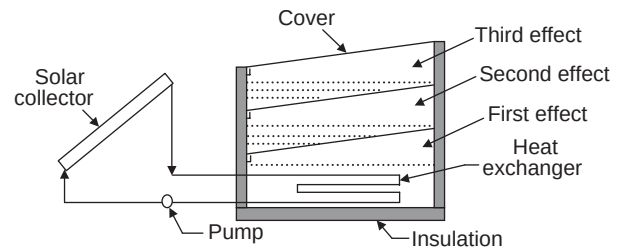


FIGURE 4 Schematic of a multi-effect solar still.

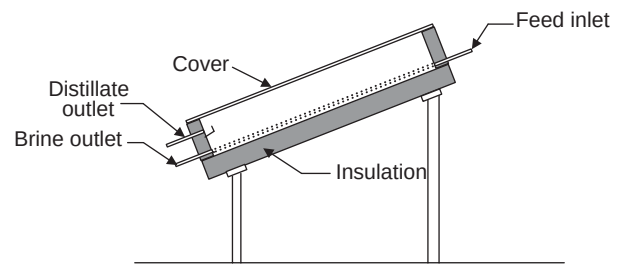


FIGURE 5 Schematic of a tilted solar still.

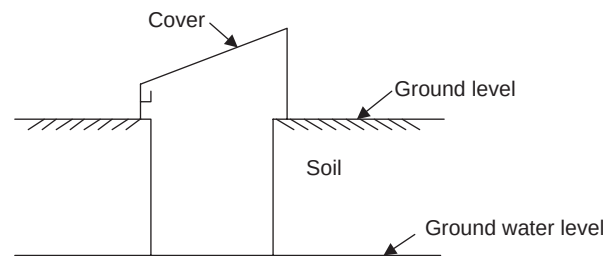


FIGURE 6 Humidity solar still.

circulated in the opposite direction and becomes humid. The humid air goes to the condenser, where the condensing vapor creates fresh water. In humid areas where there is no fresh or saline water available, fresh water can be obtained from the atmosphere. This can be done by using desiccant materials, which absorb water vapor from the atmosphere. The cooling coil of an air conditioner may be used to produce fresh water, but its temperature must be lower than the dew point of the humid air. As the humid air comes in contact with the coil, the water vapor temperature reduces below the dew point, which condenses as fresh water. For locations with very low humidity and no water resources, fresh water may be obtained from the soil moisture. At a certain depth below the earth surface, the soil remains moist. If that layer of moist soil is heated, vapor will be formed that can be condensed and collected as fresh water. Figure 6 shows a schematic of a possible arrangement.

4. STILL PERFORMANCE

Although the construction and operation of a solar still is simple, the theoretical analysis is complicated and mostly depends on experimental observations. Basic concepts of a solar still were set forth by R. V. Dunkle in 1961.

The mass transfer rate $\dot{m}$, in (Kg/m².s), can be written as

$$\dot{m} = 9.15 * 10^{-7} h_c (p_{wb} - p_{wg}).$$

The heat transfer by evaporation-condensation is

$$q_e = 9.15 * 10^{-7} h_c (p_{wb} - p_{wg}) * h_{fg},$$

where, h_c is the convection coefficient in the still, given by

$$h_c = 0.884 \left[T_b - T_g + \left(\frac{p_{wb} - p_{wg}}{2016 - p_{wb}} \right) T_b \right]^{1/3},$$

where, p_{wb} and p_{wg} are the vapor pressures, in mm Hg, of the solution in the basin at the temperature of the basin (T_b) and of water at the cover temperature (T_g). Temperatures are in degrees Kelvin and h_{fg} is the latent heat of water (J/kg).

Still efficiency, η , is defined as the ratio of the heat transfer, q_e , in the still by evaporation-condensation to the incident solar radiation, I , on the still, as given by

$$\eta = \frac{q_e}{I}.$$

Still efficiency is usually low and rarely exceeds 50%, with an average of 30 to 40%. The daily production from a solar still is about 3 to 4 l/m².

5. OPERATION AND MAINTENANCE

As the water inside the still is almost stagnant, suspended solids in the feed water may settle inside the still. To prevent this, suspended material should be removed from the feed before it enters the still. Growth of algae and other organisms may occur in the still. This can be minimized through the addition of algicides such as copper sulfate to the supply. The presence of scale leads to operating difficulties or loss of efficiency by reducing the heat transfer through the affected surfaces. Therefore, scale prevention is an integral part of the design, operation, and cost of the process. The scale is formed by the dissolved solids in raw water and from corrosion products. Scale usually contains chemical compounds of magnesium oxide, calcium carbonate, and calcium

sulfate. The formation of scale can occur where the solubility limits of these compounds are exceeded. When pure water is separated from saline water, the concentration of dissolved solids increases and eventually reaches the saturation limit of one or more salts. Although supersaturated solutions are unstable, solid salts do not necessarily deposit immediately from these solutions. When these solutions come into contact with a solid crystal, the excess dissolved solid in the supersaturated solution precipitates. Crystals of the salt in supersaturated solution and other solids particles can act as centers of crystallization. Therefore, the best way to eliminate the scale formation problem is to flush the still before it reaches saturation.

Stills are designed to operate with water in them, if they are left dry for a period of time they may suffer serious damage. It is also necessary to clean the transparent cover periodically to maintain high transmittance of solar radiation.

6. SOLAR DRYING: DEFINITION AND BACKGROUND

Drying is one of the most important post harvest steps. It enhances the storage life of the crop products, minimizes losses during storage, and saves shipping costs. The drying process is the removal of water from the wet surface of the food. In this process, heat is transferred by convection and radiation to the surface of the produce. This heat raises the temperature and evaporates the moisture from the exterior of the agricultural products, diffusing the interior moisture to the surface and replenishing the evaporated surface moisture. Vapor in the surrounding air is removed by diffusion and movement of air. This process continues until the required drying ratio is reached. So the drying process requires a mechanism for heating the ambient air, a drying unit where moisture removal takes place, and an air handling unit (in some dryers).

The ability of the air to evaporate water from the produce is known as the drying potential. The drying potential depends on both air temperature and relative humidity. Besides the air drying potential, the airflow rate is another major factor determining the drying time. To reduce the drying time, the airflow rate and temperature need to be increased while and air's relative humidity needs to be decreased.

The food type of product to be dried determines the rate of moisture transfer from the interior to the

surface. Nonhygroscopic materials can be dried to zero moisture content; however, hygroscopic materials will always have residually bound moisture. The moisture content present in the agricultural product at equilibrium conditions (when it loses moisture to the environment at the same rate at which it absorbs moisture from the ambient) is known as the equilibrium moisture content or hygroscopic equilibrium. In such a case, increasing the temperature of the drying air will be more effective than increasing the flow rate. It is clear that the process of drying is controlled by the air properties, mainly temperature and relative humidity, and by the agricultural product properties, mainly moisture content.

The main objective of drying is to supply the required heat in the most effective way to yield the best result with a minimum overall expenditure of energy. For industrial drying hot combustion gases may be allowed to pass directly through the agricultural products, but this method is not preferred because of possible contamination by the unburned fuel. Electrical air heating is preferred, but expensive. In many cases, solar energy may be the most promising source for drying, as some agricultural products, like rice, may be better suited to drying by solar than by fossil fuel systems because case hardening and other damage are less likely at low temperatures.

7. DRYING PROCESSES

Drying processes range from simple natural sun drying to large-scale industrial drying. Based on the temperature required for drying, the processes may be classified as follows:

1. *Low-temperature drying.* In this slow-drying process (usually performed at temperatures between 15 to 50°C), the air is heated to about 1 to 6°C above the ambient. The main advantages of low-temperature drying include decreased chances of overheating and low capital and running costs, but the process is slow and does not work well for some climates. For low-temperature drying, low-cost solar collectors may be used to heat the air before it enters the dryer.

2. *High-temperature drying.* In this method, air at temperatures above 50°C is used to dry the agricultural products at a fast rate. In such systems, exhaust air from the dryer may be too hot and unsaturated to be wasted, so circulation becomes necessary to improve effectiveness and reduce costs. For this kind of drying, simple solar collectors may

not be able to provide the required temperature. Concentrating collectors could be used to achieve the required temperature; however, it may be more economical to use simple solar collectors to preheat the air, and then increase the temperature of the air further using conventional fuel.

3. *Freeze drying.* In such a process, the agricultural product to be dried is first frozen, then placed inside a vacuum chamber connected to a condenser at a low temperature or a desiccant. Heat is applied to the products allowing the water to sublime, which condenses in the condenser or is absorbed by the desiccant.

4. *Osmovac drying.* This process is normally used for drying fruit juices to make a concentrated powder. It involves two processes, an osmotic process followed by vaporization under vacuum.

5. *Desiccant drying.* In this process, a desiccant is used to remove moisture from the air to increase its drying potential. The desiccant is then regenerated by solar energy. Sometimes a portion of the product itself (especially grain) might be used as the desiccant. If solar energy is used for drying, grain could be overdried during the summer in a desiccant system. At the time of harvest, a part of the overdried grain is removed from the desiccant bin and blended with wet grain to produce a mixture with the average moisture content around the required value. Low-temperature drying may be used to achieve a uniform value for this mixture (i.e., moving moisture from wet to dry agricultural products to reach the required value).

8. SOLAR DRYING PROCESSES

Using the sun's energy to dry agricultural products (known as sun drying) is the most widely used form of drying because it is cheap, easy, and convenient. Not long ago, sun drying was the only form of agricultural drying, and it is still the most common agricultural process industry. The easiest and simplest way of sun drying is to spread the products such as fruits, vegetables, cereals, tobacco, and so forth on a suitable surface under the sun and turn them regularly until sufficiently dried so that they can be stored safely. Another important application of sun drying is the evaporation of seawater to produce salt. It is obvious that sun drying would work best in a hot and dry climate with gentle breezes but would be difficult in humid areas. Although sun drying requires little capital or expertise, it has many limitations: irregular and slow drying rate, high final

moisture content especially in humid areas, and variable product quality where some of the products may be overdried and others underdried since there is no control on the drying process. Other problems include contamination with dust and insect infestation, and sudden rain storms, which can soak the products. Also, large areas are required because sunlight is not concentrated. Problems associated with sun drying can be reduced by using mechanical dryers and conventional energy sources like electricity, coal, and fossil fuels. Due to the high cost of conventional energy sources and the associated environmental problems, solar agricultural dryers have gained popularity, particularly in developing countries. Many of these countries are located in areas that receive a good amount of solar radiation.

9. SOLAR DRYERS

Solar dryers are special structures that enhance the drying power of the sun and protect the agricultural product from dust, dirt, and insects. Figure 7 shows a block diagram of an active solar drying system, consisting of solar air collectors, a drying unit or chamber, and an air handling unit. An active solar drying system can accommodate the use of a backup heat source when there is not enough solar heat available.

Solar dryers usually employ high airflow rates at low temperatures over long periods of time, whereas fossil fuel dryers usually employ lower airflow rates and higher temperatures for fast drying. Slow drying of fruits and vegetables preserves the flavor and quality of the products much better, which makes solar drying more attractive. Therefore, many commercial solar dryers have already been marketed.

Drying capacity depends on the size of the dryer and varies with the type of the produce and the amount of moisture to be removed. Placing agricultural products on racks above each other limits the area of exposure of the food surfaces to the drying air, which might result in poor drying. Underloading

will result in low efficiency because of not fully using the available capacity. The loading density can be assessed based on the following rules of thumb, which were suggested by M. A. Leon, S. Kumar, and S. C. Bhattacharya in 2002.

1. Average dryer load is about 4 kg of fresh produce per square meter of tray area.
2. Solar collector area is about 0.75 times total tray area.
3. Airflow rate is about 0.75 cubic meter per minute per meter square tray area.

There are different types of solar dryers; most of them were developed for specific products or class of products. Selection of a solar dryer represents a compromise between dryer cost, product quality, safety consideration, and installation convenience. Solar dryers may be classified into many categories depending on the mode of heating or operation. According to their operating modes, these can be classified into the following categories:

1. Passive systems: Passive systems use natural forces such as buoyancy and breeze for air flow, therefore, these systems have no external energy input except solar radiation.

a. Direct solar dryers: In such dryers, the agricultural products receive the heat directly from the sun. There are many designs for direct solar dryers, the most important of which are the following:

i. Box solar dryers: These are suitable for arid and semi-arid areas, where the relative humidity is low and the amount of incident solar radiation is high. The low initial cost and easy maintenance are the main advantages of the box dryer. Although box dryers cut the drying period by half in comparison to open sun drying, this period is still very long. Box dryers are usually rectangular, painted black on the inside, and have doors to load and unload the produce. The bottom and sides of the box may be insulated. It has transparent cover on its top. Holes are made at the base and the top of the rear vertical

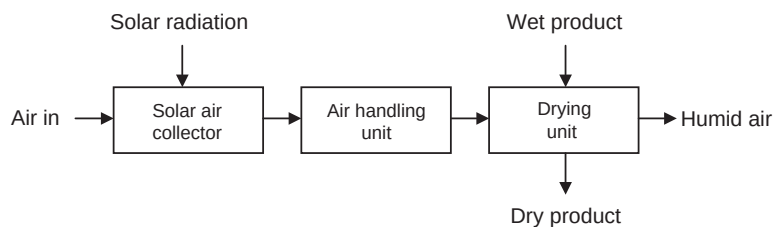


FIGURE 7 Block diagram of a solar dryer.

wall (a few centimeters below the cover) to permit movement of air by natural convection. Ambient air enters through the base holes and leaves through the top holes carrying moisture from the products. The agricultural products to be dried are kept in trays, with perforations allowing air flow through them. The trays are placed so that the product receives solar radiation directly. Heat is absorbed directly by the products themselves as well as by the internal surfaces of the drying chamber. The absorbed heat evaporates moisture from the food and induces natural convection, assisting the drying by air circulation. Figure 8 shows a schematic of a possible arrangement known as the solar box dryer. The box's overall dimensions recommend that the length is at least three times the width to minimize the shading effect. Box solar dryers can be made portable or stationary.

ii. Greenhouse solar system: A greenhouse dryer is usually made from a metallic frame covered by transparent plastic or glass with removable inner chambers, which acts as the absorber. Insulating

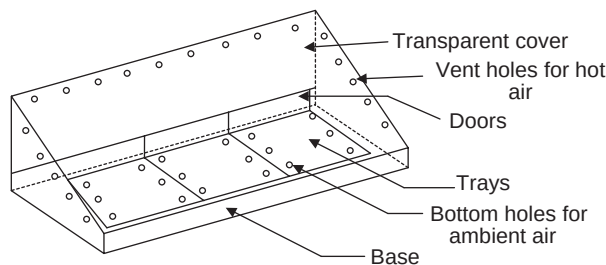


FIGURE 8 Schematic of a box type solar dryer.

panels may be used over the glazing at night to reduce heat loss, and a means of storing excess daytime heat is often provided. The system as a whole represents a combination of a solar collector and an integrated energy storage device. Figure 9 shows a greenhouse solar system with tobacco curing and grain drying modes (at the left and right, respectively), with an illustration of typical air flow patterns for solar drying and furnace drying. The incident solar radiation gets absorbed on the moveable black absorber. The useful heat is continuously removed by the slow rotation of the absorber and transferred to the agricultural products. Under such conditions, the dryer will most likely be an active one. For nongrain products, the absorber remains stationary. Such systems are usually of medium capacity.

iii. Rack type solar dryers: These dryers are mostly used for drying grapes. The dryer usually consists of 6 to 10 racks, each about 50 to 100 m long, 2.5 m high, and 1.5 m wide, with a north-south orientation lengthwise. Each rack is fitted with galvanized wire netting, spaced at about 25 cm from each other. The outside frame can be made of steel or wood. Products are placed or hung from the wire mesh. The whole structure is normally covered with iron sheets, and side curtains are sometimes used for protection from dust, rain, and excessive sunshine. Racks are usually constructed 9 to 12 m apart to ensure that the agricultural products are exposed to sunlight during the morning and afternoon hours. The racks are preferably placed in an exposed area, clear of any obstruction to permit free air flow.

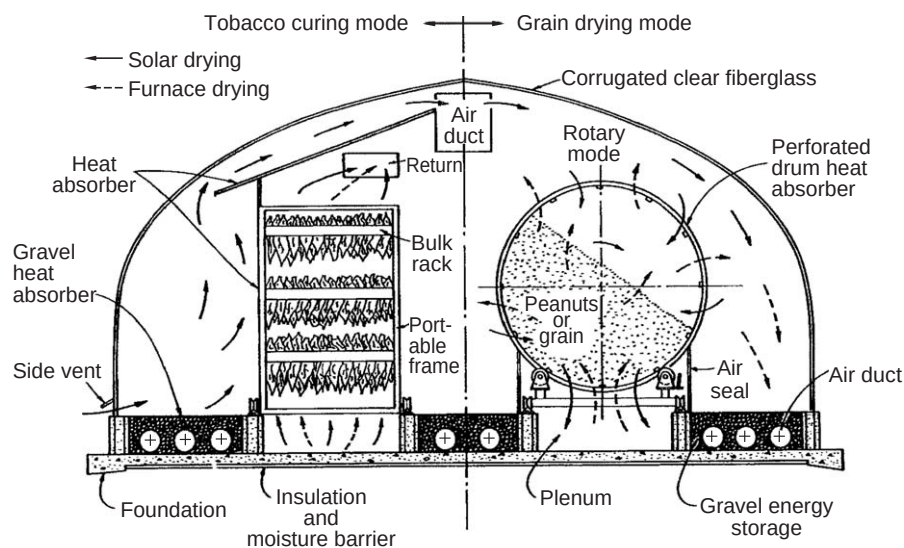


FIGURE 9 Greenhouse solar system.

b. Indirect solar dryers: This mode of drying is important especially when the products must be dried under certain specifications of temperature, humidity, and drying rate. In such dryers, the products are placed inside an opaque drying chamber, and solar collectors, usually air type, provide the hot air. Dryers of this type consist of solar air collectors and a drying chamber. The design of such dryers is the same as the conventional fuel type—the only difference is that the heat source is completely or partially solar energy. Some indirect solar dryers are described next:

i. Solar cabinet dryers: The air is heated in the collector and moves upward toward the bed due to buoyancy forces. The products to be dried are laid on the bed, which usually consists of many trays, placed close to the exit section of the collector. As the air moves across the bed, it picks up moisture from the products. Figure 10 shows one possible design of such systems.

ii. Geodesic dome solar fruit dryers: The dome shape (Fig. 11) of these greenhouse dryers

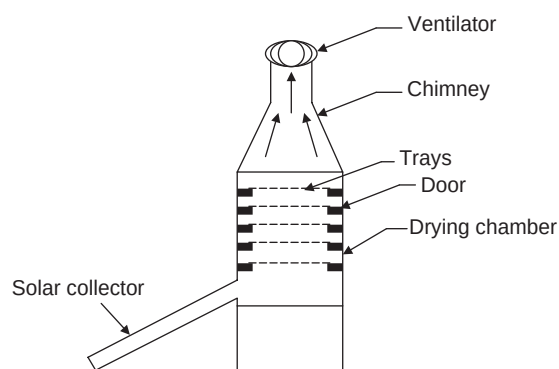


FIGURE 10 Schematic of a cabinet type solar dryer.



FIGURE 11 Structure of a geodesic dome dryer.

makes them perform better than the conventional gable and semicylindrical shapes because they are able to collect more energy from the sun. The dryers are formed from a dome structure covered from the outside by a transparent cover and from the inside by a black plastic absorber. The transparent cover and the black absorber are separated by a suitable distance, about 1 foot. The drying trays are placed inside the inner dome. Air between the transparent cover and the black absorber is heated up by solar radiation and passed through the drying trays before it leaves the dome. The air flow can be by forced convection (Fig. 12A) or by natural convection (Fig. 12B).

c. Mixed type dryers: In mixed type of dryers, the products are heated by the hot air from the solar collectors and by direct solar radiation absorbed by them. The sides and the top of the drying chamber can be made of a transparent material, where the products absorb the solar radiation directly. Figure 10 is one possible arrangement showing the front and side walls of drying chamber as transparent.

2. Active systems: An electricity-driven blower or fan is used to circulate the air in active systems. These are further classified as follows:

a. Direct solar dryers: These consist of a drying chamber with a transparent cover and a blower. The solar radiation passing through the transparent cover heats the air. The blower forces the air to pass through the products where it picks up the moisture. Figure 8 shows one possible arrangement with a blower or a fan fixed at one side to circulate the air.

b. Indirect solar dryers: In such dryers, the air heated in the solar collectors is ducted to the drying chamber, where it evaporates and removes the moisture. Some of the most important designs of

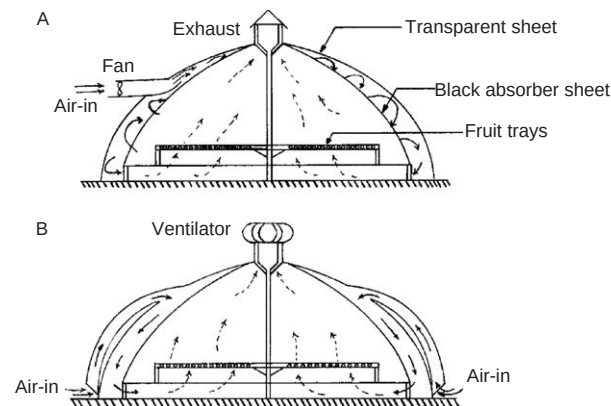


FIGURE 12 Modes of airflow in a geodesic dome dryer. (A) Forced air configuration. (B) Natural air configuration.

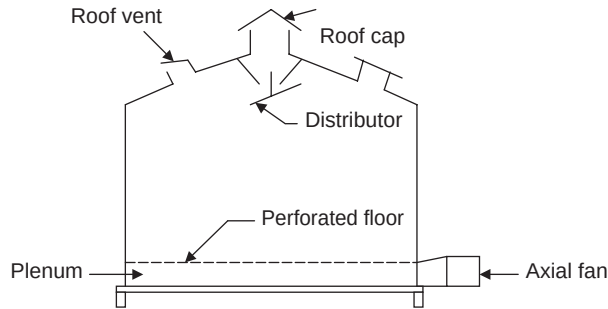


FIGURE 13 Schematic of a bin dryer.

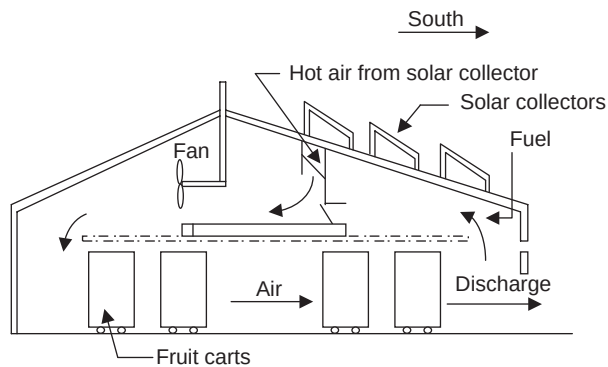


FIGURE 14 Schematic of an industrial dryer.

this type are the following:

i. Bin-type grain dryers: In these dryers, the bin is filled to the top and the hot air from the solar collector is blown from the bottom. Figure 13 shows a schematic diagram of this type.

ii. Shelf-type dryers: The drying chamber consists of many holding trays to increase the capacity of the dryer. The dryer design is similar to that in Fig. 10, but the air is circulated with the help of a blower.

iii. Hybrid or industrial dryers: Such systems use a backup source of conventional energy when solar energy is not sufficient or during off-sun hours. These are usually of a continuous process type and have large capacities. Figure 14 shows a schematic of such dryers.

c. Mixed-type dryers: These are similar to passive dryers, but air is circulated with the help of blowers or fans.

10. DRYER PERFORMANCE

The amount of heat utilized in evaporating the water, q_e , can be calculated as

$$q_e = \dot{m}_e h_{fg},$$

where h_{fg} is the latent heat of evaporation, and $\dot{m}_e$ is the amount of water evaporated in the dryer per unit area perpendicular to the direction of the flow:

$$\begin{aligned} \text{Dryer overall thermal efficiency, } \eta_d(\%) &= \frac{\text{output}}{\text{net input}} * 100 \\ &= \frac{q_e}{q_{\text{input}}}. \end{aligned}$$

If a solar collector of efficiency η_c is used to supply the hot air, the overall efficiency of the system will be

$$\eta_{\text{overall}} = \eta_d \eta_c.$$

In the case of active solar dryers, the energy consumed by the blower or fan is accounted for in the input. Also in the case of hybrid dryers, the amount of conventional energy consumed is added in the solar input.

SEE ALSO THE FOLLOWING ARTICLES

Solar Cookers • Solar Cooling, Dehumidification, and Air-Conditioning • Solar Detoxification and Disinfection • Solar Fuels and Materials • Solar Heat Pumps • Solar Thermal Energy, Industrial Heat Applications • Solar Thermal Power Generation • Solar Water Desalination

Further Reading

- Brace Research Institute (1975). A survey of agricultural dryers, Technical report T99. Faculty of Engineering, McGill University, Quebec, Canada.
- Delyannis, E., and Belessiotis, V. (2001). Solar energy and desalination. In "Advances in Solar Energy: An Annual Review of Research and Development." (D. Y. Goswami, Ed.), Vol. 14, pp. 287–330. American Solar Energy Society, Boulder, CO.
- Dunkle, R. V. (1961). Solar water distillation: The roof type still and a multiple effect diffusion still. International Conference of Heat Transfer, University of Colorado, Publication No. 108, pp. 895–902.
- Ekechukwu, O. V., and Norton, B. (1999). Review of solar-energy drying systems III: Low temperature air-heating solar collectors for crop drying applications. *Energy Conversion Management* 40, 657–667.
- Goswami, D. Y., Kreith, F., and Kreider, J. F. (2000). "Principles of Solar Engineering," 2nd ed. Taylor & Francis, Philadelphia.
- Kalogirou, S. (1997). Survey of solar desalination systems and system selection. *Energy* 22(1), 69–81.
- Leon, M. A., Kumar, S., and Bhattacharya, S. C. (2002). A comprehensive procedure for performance evaluation of solar food dryers. *Renewable Sustainable Energy Rev.* 6, 367–393.

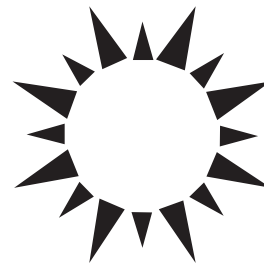
- Lior, N. (1986). "Desalination and Water Purification, 2. Measurements and Control in Water Desalination." Elsevier, Amsterdam.
- Pangavhane, D. R., and Sawhney, R. L. (2002). Review of research and development work on solar dryers for grape drying. *Energy Conversion and Management* 43, 45–61.
- Sodha, M. S., Bansal, N. K., Kumar, A., Bansal, P. K., and Malik, M. A. S. (1987). "Solar Crop Drying." CRC Press, Boca Raton, FL.
- Sodha, M. S., and Chandra, R. (1994). Solar drying systems and their testing procedures: A Review. *Energy Conversion and Management* 35(3), 219–267.
- Spiegler, K. S. (1977). "Salt Water Purification." 2nd ed. Plenum Press, New York.
- United Nations (1970). "Solar Distillation as a Means of Meeting Small-Scale Water Demands." New York.

Solar Energy, History of

JOHN PERLIN

Rahus Institute

Martinez, California, United States



-
1. Solar Architecture and Greenhouses
 2. Solar Thermal Power
 3. Solar Hot Water Heating
 4. Photovoltaics

Glossary

solar architecture The conception, design, and construction of buildings and communities so as to utilize incoming solar radiation.

solar cell A semiconductor material often made of specially treated silicon that converts sunlight directly into electricity.

solar greenhouses Structures used to cultivate plants that are designed and built to utilize solar energy for both heating and lighting.

solar thermal power Systems that capture energy from solar radiation, transform it into heat, and then use it for a variety of applications.

The ancient Greeks began using solar energy 2500 years ago. Since then, people have used the Sun for their benefit.

1. SOLAR ARCHITECTURE AND GREENHOUSES

1.1 Ancient Greece

During the fifth-century BC, the Greeks faced severe fuel shortages. Fortunately, an alternative source of energy was available—the sun (Fig. 1). Archaeological evidence shows that a standard house plan evolved during the fifth century so that every house, whether rural or urban, could make maximum use of the sun's warm rays during winter (Fig. 2). Those living in ancient Greece confirm what archaeologists have found. Aristotle noted that builders made sure

to shelter the north side of the house to keep out the cold winter winds. Socrates, who lived in a solar-heated house, observed, “In houses that look toward the south, the sun penetrates the portico in winter,” which keeps the house heated. The great playwright Aeschylus went so far as to assert that only primitives and barbarians “lacked knowledge of houses turned to face the winter sun, dwelling beneath the ground like swarming ants in sunless caves.” The ancient Greeks planned whole cities in Greece and Asia Minor, such as Colophon, Olynthus, and Priene, to allow every homeowner access to sunlight during winter to warm their homes. By designing the streets in a checkerboard pattern running east–west and north–south, every home faced south, permitting the winter sun to flow into all houses.

1.2 Ancient Rome

Fuel consumption in ancient Rome was even more profligate than in classical Greece. In architecture, the Romans remedied the problem in the same fashion as did the Greeks. Vitruvius, the preeminent Roman architectural writer of the first century BC, advised builders in the Italian peninsula, “Buildings should be thoroughly shut in rather than exposed toward the north, and the main portion should face the warmer south side.” Varro, a contemporary of Vitruvius, verified that most houses of at least the Roman upper class followed Vitruvius's advice, stating, “What men of our day aim at is to have their winter rooms face the falling sun [southwest].” The Romans improved on Greek solar architecture by covering south-facing windows with clear materials such as mica or glass.

Such clear coverings let in the light. When the light hit the masonry it converted into heat. The longer heat waves could not easily escape through the glass and the area inside heated to higher temperatures than those of previously unglazed structures. The Romans called these glass-covered structures *heliocamini*,

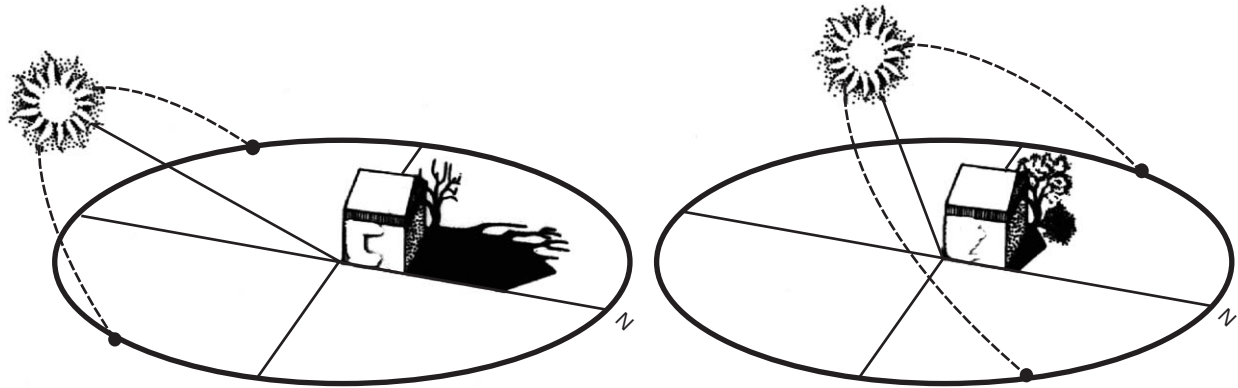


FIGURE 1 In the Northern Hemisphere, the low winter sun rises relative to Earth in the southeast, by noon it is in the south, and it sets in the southwest. During summer, the sun rises in the northeast, by noon it is directly south and very high in the sky, and it sets in the northwest. From Perlin/Butti Solar Archives.

which means “sun furnaces.” The Romans used the term to describe their south-facing rooms because they became much hotter in winter than similarly oriented Greek homes, which lacked such coverings.

1.3 Ancient Roman and Renaissance Solar Greenhouses

The ancient Romans not only used window coverings to hold in solar heat for their homes but also relied on such solar heat traps for horticulture. Plants would mature quicker and produce fruits and vegetables out of season and exotic plants from hotter climates could be cultivated at home. With the fall of the Roman Empire, so too came the collapse of glass for either buildings or greenhouses. Only

with the revival of trade during the 16th century was there renewed interest in growing in solar-heated greenhouses exotics brought back from the newly discovered lands of the East Indies and the Americas. Trade also created expendable incomes that allowed the freedom to take up such genteel pursuits as horticulture once more. Nineteenth-century architects such as Humphrey Repton brought the sunlit ambiance of the greenhouse into the home by attaching it onto the south side of a living room or library. On sunny winter days, the doors separating the greenhouse and the house were opened to allow the moist, sun-warmed air to freely circulate in the otherwise gloomy, chilly rooms. At night, the doors were shut to keep in as much solar heat as possible. By the late 1800s, the country gentry had become so enamored of attached greenhouses that they became an important architectural fixture of rural estates.

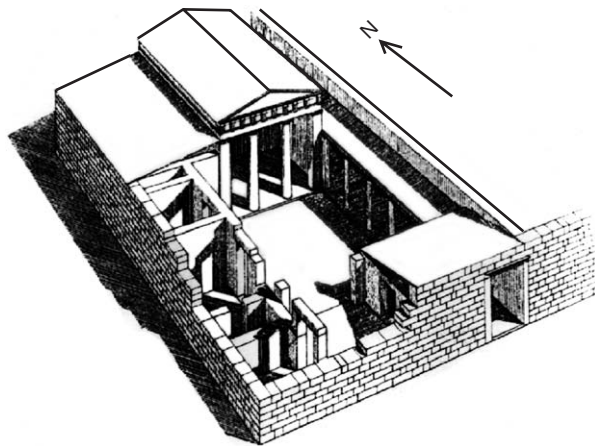


FIGURE 2 The Greeks took advantage of their knowledge of the location of the sun during the year to heat and cool their buildings. They opened their buildings to the south and protected them with a covered porch. The strategy allowed winter sunlight to pour into the house, whereas the roof shaded the house from the high summer sun. From Perlin/Butti Solar Archives.

1.4 Indigenous America

American solar architecture began with its indigenous heritage. Acoma, built by the Pueblo Indians in the 12th century and continuously inhabited since then, serves as an excellent example of their sensitivity to building with the climate in mind. The “sky city,” built atop a plateau, consists of three long rows of dwelling units running east to west. Each dwelling unit has two or three tiers stacked one behind the other to allow each one full exposure to the winter sun. Most doors and windows open to the south to catch the warm solar rays of winter. The walls are built of adobe. The sun strikes these heat-absorbing south walls more directly in winter than during summer. The sun’s heat travels through the adobe slow enough to reach the interior as night

falls, heating the house through the night. Insulating the ceiling with straw keeps the horizontal roof, vulnerable to the rays of the high summer sun, from allowing too much heat to enter the house.

1.5 Early Euro-Americans

Spaniards who settled in the American southwest often built to take advantage of the winter sun: They aligned their homes east to west so the main portion of the house faced south. Shutters outside the windows were closed at night to help keep the solar heat that had flowed in all day from escaping during wintertime.

Settlers in New England considered the climate when they built their homes. They often chose “saltbox” houses that faced toward the winter sun and away from the cold winds of winter. These structures had two south-facing windowed stories in front where most of the rooms were placed and only one story at the rear of the building. The long roof sloped steeply down from the high front to the lower back side, providing protection from the winter winds. Many saltbox houses had a lattice overhang protruding from the south facade above the doors and windows. Deciduous vines growing over the overhang afforded shade in summer but dropped their leaves in winter, allowing sunlight to pass through and penetrate the house.

1.6 Post-World War I Europe and America

Ameliorating the terrible slums that blighted European cities during the industrial revolution sparked a renewed interest in building with the sun. As one English city planner bent on housing reform urged in the first decade of the 20th century, “Every house should have its face turned to the sun, whence comes light, sweetness, and health.” Young architects on the Continent agreed, declaring that their profession should embrace functionalism over aesthetics. To many, functionality meant designing houses that satisfied the basic needs of those living in them. In the relatively chilly European climate, this dictum meant providing dwellers with a warm haven. Therefore, Hannes Meyer, the leader of functional architecture in Germany, argued that an architect wishing to build a useful house must consider its body “to be a storage cell for the heat of the sun.” Hence, “maximal use of the sun,” as far as Meyer was concerned, was “the best plan for construction.”

Meyer, who became the director of the highly famous and influential Bauhaus architectural school, told fellow architects that to achieve optimum sunshine in buildings, they must conduct or at least be privy to scientific research into the sun’s year-round movement relative to the earth. Germany led a renaissance in solar building during the 1930s that spread throughout Europe only to be stamped out by the Nazis, who condemned functional architecture as Jewish. When they came to power, a good number of German architects who designed solar buildings fled, many to the United States. George Fred Keck, a Chicago architect, befriended some of these expatriates and through their influence began designing homes in the Chicago area according to solar building principles: expansive south-facing glass to trap the winter sun; long overhangs to shade the house in summer; minimal east–west exposure to prevent overheating in summer and fall; and the placement of secondary rooms, garages, and storage corridors on the north side to help insulate the living quarters from the cold north winds. Keck had a knack for publicity and called the houses he designed “solar homes.” By the mid-1940s, Keck’s work caught the attention of the national media. *House Beautiful*, *Reader’s Digest*, and *Ladies Home Journal* featured his work. Fuel rationing during the war inclined the American public toward valuing the energy-saving features of solar homes. When war ended, the building market exploded. With the wartime-conservation ethic still imbued in most people’s minds, many manufacturers in the prefabricated home industry adopted solar design features for leverage in this highly competitive market.

1.7 Post-World War II America

Studies of houses with large amounts of glass on the south side demonstrated that they experience much greater temperature swings than ordinary homes, warming quickly during the day due to accumulating the sun’s heat and cooling off rapidly at night due to large heat losses through the windows. Tucson Architect Arthur Brown eliminated the unwanted fluctuations in temperature by running a thick black wall between the sun porch next to the south-facing glass and the living quarters. During sunny winter days, sunlight struck the black wall and the concrete absorbed the solar heat. Estimating that heat moves through the wall at 1 in. per hour, Brown made the wall 8 in. thick. As evening approached, the solar heat soaked up by the wall began to radiate into the rooms and continued to do so during the night. By

morning, all the solar heat in the wall had dissipated into the living areas, leaving the wall totally cooled down and ready for another heating cycle.

Brown also designed the world's first solar-heated public building in 1948. Since students attended classes between 9 AM and 3 PM, Brown did not have to worry about storing solar energy since no one needed heat at night or in the very early morning. To keep costs down, he used the roofing supports to separate the space between the ceiling and the roof into channels. In this way, the roof served as the solar collector. The school's fan, which had forced gas-heated air through the building, now pushed air through the channels. The sun's energy heated the air and a second fan distributed that heat into the classrooms when needed. The solar system provided the school with 86% of its heat. However, in 1958, when the school district decided to expand the campus, the authorities chose a gas-fired furnace due to cheap energy costs. Cheap energy running highly reliable and easy to use heating and cooling systems led to an almost universal disinterest in solar design. It was only in the 1970s, with the Arab oil embargo and skyrocketing fuel prices, that people became interested in solar architecture again.

2. SOLAR THERMAL POWER

2.1 Ancient Chinese, Greeks, and Romans

Archaeological finds and ancient literature show that solar concave reflectors and convex lenses have a long history. Convex lenses predate concave reflectors by more than 1000 years. People crafted the first convex lenses out of rock crystal. Ancient Chinese and Roman doctors relied on convex lenses to cauterize wounds. The first mention of solar concave reflectors appeared in the third-century BC writings of Euclid. The ancient Chinese and Greeks used solar concave reflectors to light fires for domestic cooking.

2.2 Medieval and Renaissance Europe

Latin translation of Arab writer Al-Hazan introduced Europeans to ancient Greek works on solar concave reflectors and convex lenses. Between the 13th and 16th centuries, many Europeans suggested burning whole armies with huge solar reflectors, but no one could build reflectors or lenses large enough for martial purposes. However, Europeans did use reflectors to ignite fires for cooking meals, to distill perfumes, and to weld metal statuary.

2.3 19th- and Early 20th-Century America, Europe, and Africa

Many in 19th-century Europe worried that the prodigious consumption of coal by industry would eventually lead to coal shortages. Among the concerned, Augustine Mouchot thought that the sun's heat could replace coal to run Europe's many steam engines. Mouchot's work with mirror technology led him to develop the world's first sun motor in the 1870s, which produced sufficient steam to drive machinery (Fig. 3). One journalist described it as a "mammoth lamp shade, with its concavity directed skyward." After 1 year of testing Mouchot's sun machine, the French government concluded,

In France, as well as in other temperate regions, the amount of solar radiation is too weak for us to hope to apply it for industrial purpose. Though the results obtained in our temperate and variable climate are not very encouraging, they could be much more so in dry and hot regions where the difficulty of obtaining other fuel adds to the value of solar technologies. [Hence] in certain cases, these solar apparatuses could be called upon to provide useful work.

An American inventor, Frank Shuman, took his cue from the French report. In 1911, he chose to locate his commercial solar motor in Egypt, where land and labor were cheap, sun was plentiful, and coal was costly. To increase the amount of heat generated by the solar motor, glass-covered pipes were cradled at the focus of a low-lying trough-like reflector. A field of five rows of reflectors was laid out in the Egyptian desert. The solar plant also boasted a storage system that collected excess sun-warmed water in a large insulated tank for use at night and during inclement weather, in contrast to the solar motor designed by Mouchot, which went "into operation only when the sun [came] out from behind a cloud and [went] out of action the instant the sun disappears again." For industrialists, Shuman's solution eliminated a major obstacle to solar's appeal. Shuman's solar plant proved in Egypt to be more economical than a coal-fired plant. Engineers recognized that this plant demonstrated that "solar power was quite within the range of practical matters." However, the hopes and dreams of solar engineering disintegrated with the outbreak of World War I. The staff of the Egyptian plant had to leave for war-related work in their respective homelands. Shuman, the driving force behind large-scale solar developments, died before the war ended. After the war, the world turned to oil to replace coal. Oil and gas reserves were found in sunny and previously

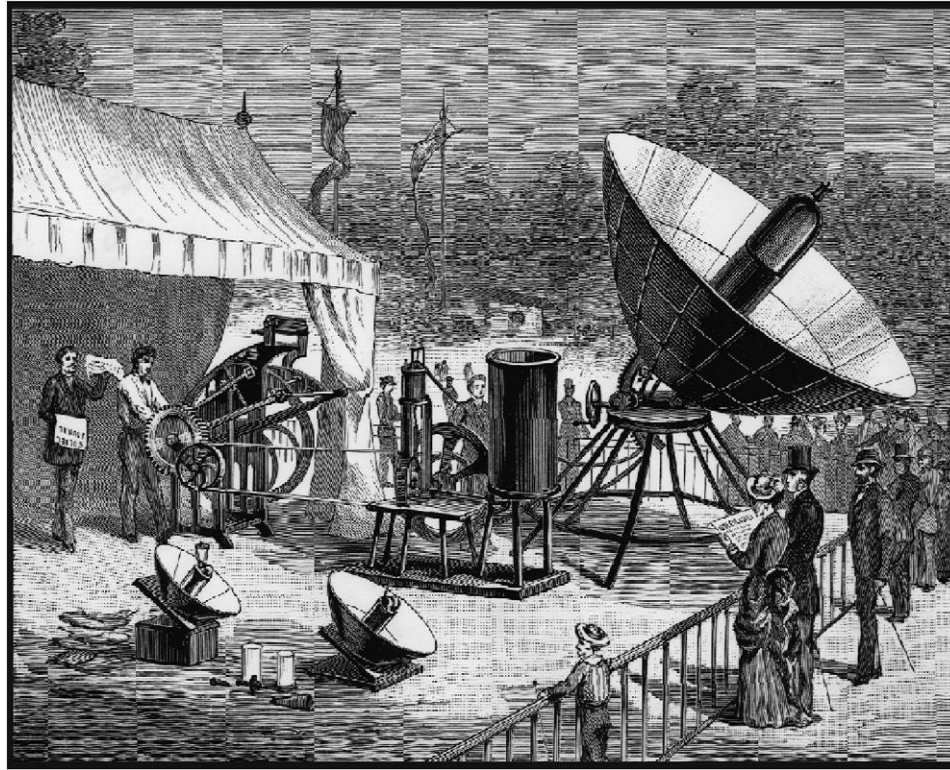


FIGURE 3 A solar-powered machine built by Mouchot's protégé, Pifre, runs a printing press. From Perlin/Butti Solar Archives.

sun-short places that had been targeted as prime locations for solar plants. With oil and gas selling at near-giveaway prices, scientists, government officials, and business people became complacent about the world's energy situation and interest in sun power came to an abrupt end.

3. SOLAR HOT WATER HEATING

3.1 Solar Hot Box

Horace de Saussure, a noted Swiss naturalist, observed in the 1760s, "It is a known fact, and a fact that has probably been known for a long time, that a room, a carriage, or any other place is hotter when the rays of the sun pass through glass." To determine the effectiveness of trapping heat with glass covers, de Saussure built a rectangular box out of half-inch pine, insulated the inside, covered the top with glass, and placed two smaller glass-covered boxes inside. When exposed to the sun, the bottom box heated to 228°F (109°C), or 16°F (9°C) above the boiling point of water. de Saussure was unsure how the sun heated the glass boxes. Today, we can

better explain what happened. Sunshine penetrated the glass covers. The black inner lining absorbed the sunlight and converted it into heat. Although clear glass allows the rays of the sun to easily enter through it, it prevents heat from doing the same. As the glass trapped the solar heat in the box, it heated up. Its inventor realized that someday the hot box might have important practical applications because "it is quite small, inexpensive, and easy to make." Indeed, the hot box became the prototype for the solar collectors that have provided sun-heated water to millions since 1891.

3.2 The First Solar Water Heaters

In the 19th century, there was no easy way to heat water. People generally used a cook stove for this purpose. Wood had to be chopped or heavy hods of coal lifted, and then the fuel had to be kindled and the fire periodically stoked. In cities, the wealthier heated their water with gas manufactured from coal. However, the fuel did not burn clean and the heater had to be lit each time someone wanted to heat water. If someone forgot to extinguish the flame, the

tank would blow up. To add to the problem of heating water, in many areas wood or coal or gas cost a lot and many times could not be easily obtained. To circumvent these problems, many handy farmers, prospectors, or other outdoorsmen devised a much safer, easier, and cheaper way to heat water: They placed a metal water tank painted black into sunlight to absorb as much solar energy as possible. These were the first solar water heaters on record. The downside was that even on clear, hot days it usually took from morning to early afternoon for the water to get hot. As soon as the sun went down, the tanks rapidly lost heat because they had no protection from the night air.

3.3 The First Commercial Solar Water Heater

The shortcomings of the bare tank solar water heaters came to the attention of Clarence Kemp, who sold the latest home heating equipment in Baltimore, Maryland. In 1891, Kemp patented a way to combine the old practice of exposing metal tanks to the sun with the scientific principle of the hot box, thereby increasing the tanks' capability to collect and retain solar heat. He called his new solar water heater the Climax, the world's first commercial solar

water heater (Fig. 4). Kemp originally marketed his invention to eastern gentlemen whose wives had left with their maids to summer resorts, leaving their husbands to fend for themselves. The solar water heater, Kemp advertised, would simplify housekeeping duties for this class of men already burdened by their wives and domestic staffs' absence and unaccustomed to such work as lighting the gas furnace or stove to heat water. In California and other temperate states, having greater amounts of sunshine throughout the year but also higher fuel costs (compared to states such as Maryland) made it essential for residents to take their solar assets seriously and not waste them (Fig. 5). The Climax sold well in such areas. Sixteen hundred had been installed in homes throughout southern California by 1900, including those installed in Pasadena, where one-third of the households heated their water with the sun by 1897.

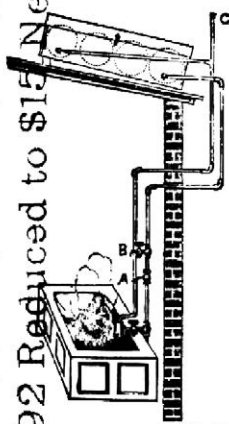
3.4 The Flat Plate Collector

From the turn of the century to 1911, more than a dozen inventors filed patents that improved on the Climax. However, none changed the fact that the heating unit and the storage unit were one and the same and both laid exposed to the weather and

Climax Solar-Water Heater
UTILIZING ONE OF NATURE'S GENEROUS FORCES

THE SUN'S HEAT { Stored up in Hot Water for Baths,
Domestic and other Purposes.

Price Of No. 1 Heater for
1892 Reduced to \$15 Net



GIVES HOT WATER at all HOURS
OF THE DAY AND NIGHT.

NO DELAY.

FLows INSTANTLY.

NO CARE. NO WORRY.

ALWAYS CHARGED.

ALWAYS READY.

THE WATER AT TIMES
ALMOST BOILS.

Price, No. 1, \$25.00

This Size will Supply sufficient
for 3 to 8 Baths.

CLARENCE M. KEMP, BALTIMORE, MD.

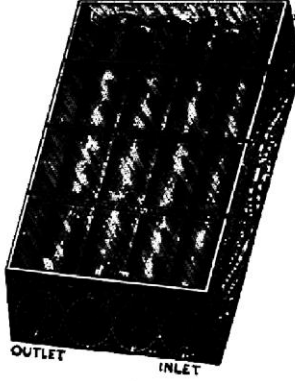


FIGURE 4 Advertisement for the world's first commercial solar water heater. From Perlin/Butti Solar Archives.



FIGURE 5 Installation of tanks in glass-covered box, Hollywood, California. From Perlin/Butti Solar Archives.

the cold night air. Hence, water heated by the sun the night before never stayed hot enough to do the wash the next morning or to heat the bath. In 1909, William J. Bailey patented a solar water heater that revolutionized the business. He separated the solar water heater into two parts: a heating element exposed to the sun and an insulated storage unit tucked away in the house so families could have sun-heated water day and night. The heating element consisted of pipes attached to a black-painted metal sheet placed in a glass-covered box. Because the water to be heated passed through narrow pipes rather than sitting in a large tank, Bailey reduced the volume of water exposed to the sun at any single moment and therefore the water heated faster. The heated water, which was lighter than the incoming cold water, naturally and immediately rose through the pipes to an insulated storage tank, where the water was kept warm for use both day and night. He also connected the storage tank to the home's furnace, guaranteeing hot water even after several rainy days. Providing hotter water for longer periods put Bailey's solar hot water heater, called the Day and Night (Fig. 6), at a great advantage over the competition. Soon, the Climax company went out of business. From 1909, when Bailey started

his business, through 1918, his company sold more than 4000 Day and Night solar hot water heaters (Fig. 7).

3.5 The Demise of the Solar Water Heater in California and Its Success in Florida

The discovery of huge amounts of natural gas in the Los Angeles basin during the 1920s and 1930s killed the local solar water heater industry. Rather than lose money from the energy changes, Bailey applied the innovations he had made in solar to develop thermostatically controlled gas water heaters. His Day and Night gas water heater made him his second fortune. He also sold the patent rights of the Day and Night solar water heater to a Florida firm. A building boom in Florida during the 1920s had tripled the number of houses, but just as in California before the great oil strikes, people had to pay a high price to heat water. The high cost of energy combined with the tropical climate and the great growth in housing stock created a major business for those selling solar water heaters. By 1941, more than half the population heated water with the sun. Declining electric

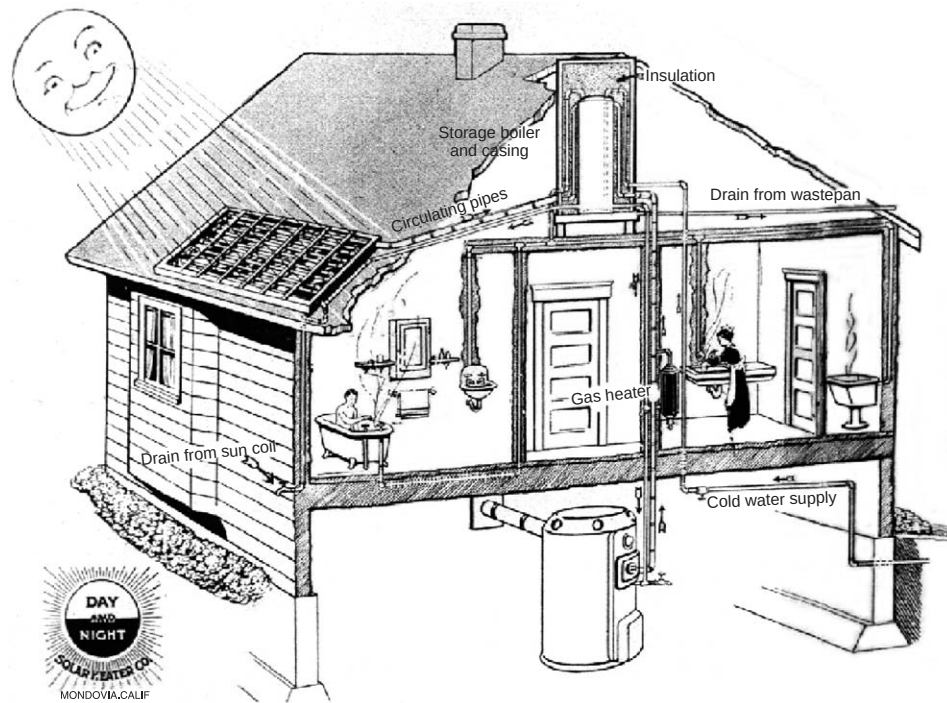


FIGURE 6 Schematic of the first flat plate solar collector with solar heat collection and solar storage separated. From Perlin/Butti Solar Archives.



FIGURE 7 An early flat plate collector installation. From Perlin/Butti Solar Archives.

rates after World War II, in association with an aggressive campaign by Florida Power and Light to increase electrical consumption by offering electric

water heaters at bargain prices, brought Florida's once flourishing solar water heater industry to a screeching halt.

3.6 Solar Water Heaters in Japan

Unlike in the United States during the post-World War II years, Japan lacked cheap and abundant energy to supply hot water on demand. Rice farmers in particular yearned for a hot bath after working long hours in the hot, humid paddy fields. To heat water, however, they had to burn rice straw, which they could have otherwise used to feed their cattle or fertilize the earth. A Japanese company began marketing a simple solar water heater consisting of a basin with its top covered by glass, and more than 100,000 were in use by the 1960s. People living in towns and cities bought either a plastic solar water heater that resembled an inflated air mattress with a clear plastic canopy or a more expensive but longer lasting model that resembled the old Climax solar water heaters—cylindrically shaped metal water tanks placed in a glass-covered box. Approximately 4 million of these solar water heaters had been sold by 1969. The advent of huge oil tankers in the 1960s allowed Japan access to new oil fields in the Middle East, supplying it with cheap, abundant fuel. As had happened in California and Florida, the solar water heater industry collapsed, but not for long. The oil embargo of 1973 and the subsequent dramatic increase in the price of petroleum revived the local solar water heater industry. Annual sales of more than 100,000 units continued to hold steady from 1973 until the second oil shock of 1979. Sales then increased to almost 500,000 that year and increased further to nearly 1 million the following year. By this time, the Japanese favored solar water heaters that closely resembled the type introduced in California in 1909 by William J. Bailey with the heating and storage units separated. As the price of oil began to stabilize in 1985 and then decline sharply in subsequent years, so did the sales of solar water heaters; however, the Japanese still purchase approximately 250,000 each year. Today, more than 10 million Japanese households heat water using the sun.

3.7 Solar Water Heating in Australia

From the 1950s to the early 1970s, a few thousand Australians relied on the sun to heat water. The numbers grew as a consequence of the major spikes in oil prices in 1973 and 1979. Interestingly, the purchase of solar water heaters during these years varied from state to state. Whereas 40–50% of those living in Australia's Northern Territory heated water using the sun, approximately 15% in Western Australia did so and less than 5% did so in the more populated eastern states. The major difference had

more to do with the cost of electricity than the amount of sun available. People in the Northern Territory and Western Australia bought electricity generated by imported and increasingly costly petroleum, whereas for those in the eastern states of New South Wales, Queensland, and Victoria, electricity was produced by locally mined and very cheap coal. In the late 1980s, the Australian solar water heater market began to stagnate. Pipelines bringing newly discovered natural gas to previously fuel-short regions, such as the Northern Territory and Western Australia, eliminated growth in these once fertile markets for solar water heaters. Exports now account for more than 50% of the sales of Solahart, Australia's leading manufacturer of solar water heaters.

3.8 Solar Hot Water Heaters in Israel

Unlike the United States and much of Europe, Israel, like Japan, found itself without sufficient fuel supplies in the early 1950s. The power situation became so bleak in the early days of the Jewish state that the government had to forbid heating water between 10 PM and 6 AM. Despite mandatory domestic rationing of electricity, power shortages worsened, causing pumping stations to fail and threatening factory closures. A special committee impaneled by the government could only suggest the purchase of more centralized generators to overcome the problem. This conclusion raised the ire of Israeli engineer Levi Yissar, who complained, "How about an already existing energy source which our country has plenty of—the sun. Surely we need to change from electrical energy to solar energy, at least to heat our water." Yissar put his money where his mouth was, becoming Israel's first manufacturer of solar water heaters. His design closely resembled the type introduced in California in the early 20th century with heating and storage separated. By 1967, approximately 1 in 20 households heated water with the sun. However, cheap oil from Iran in the late 1960s as well from oil fields captured during the Six Day War drastically reduced the price of electricity and the number of people purchasing solar water heaters. Israeli success in the Yom Kippur War resulted in the infamous oil boycott of 1973. The Israelis responded by mass purchasing of solar water heaters. By 1983, 60% of the population heated water using the sun. When the price of oil dropped in the mid-1980s, the Israeli government did not want people backsliding in their energy habits as happened throughout the rest of the world. It therefore

required its inhabitants to heat water using the sun. Today, more than 90% of Israeli households own solar water heaters.

3.9 Solar Water Heating for Swimming Pools

Solar swimming pool water heaters rank as the most successful but least heralded commercial solar application. The use of solar energy for pool heating and the equipment and needs of pool owners make a perfect match. The storage unit for the solar heated water already exists—the swimming pool. The pump needed to push water through the solar collectors must be purchased irrespective of the technology used to heat the water. The pool owner merely has to purchase the solar collectors. Since those using the pool generally want the temperature of the pool to be no higher than 80°F (27°C), the solar collectors do not require a costly glass cover or expensive metal sheeting and piping. Climax solar water heaters kept pools warm in the 1890s. Later, collectors resembling those designed by William J. Bailey were put to use for this application. In the 1970s, American Freeman Ford developed low-cost plastic to act as the solar collector. Exposed to the sun, water would pass through narrow ducts in the plastic and carry enough heat to warm the pool. The outdoor swimming season harmoniously coincides with the maximum output of the solar collectors. Even with other forms of energy selling very cheaply, the pool owner who buys a solar unit will start to save money very quickly. In the United States alone, solar swimming pool heaters have produced the energy output equivalent of 10 nuclear power plants.

4. PHOTOVOLTAICS

All the methods discussed in the previous sections change solar energy into heat. In contrast, photovoltaics produces electricity directly from sunlight.

4.1 The First Solid-State Solar Cell

The direct ancestor of solar cells currently in use got its start in the late 19th century with the construction of the world's first seamless communication network—transoceanic telegraph cables. While laying them under water to permit instantaneous commu-

nications between continents, engineers experimented with selenium for detecting flaws in the wires as they were submerged. Those working with selenium discovered early on that the material's performance depended on the amount of sunlight falling on it. The influence of sunlight on selenium aroused the interest of scientists throughout Europe, including William Grylls Adams and his student Richard Evans Day. During one of their experiments with selenium, they witnessed something completely new: Light could cause a solid material to generate electricity.

However, with the science of the day not yet sure if atoms were real, scientists could not explain why selenium produced electricity when exposed to light. Most scientists of the day therefore scoffed at Adams and Day's work. It took the discovery and study of electrons and the discovery and acceptance that light contains packets of energy called photons for the field of photovoltaics to gain acceptance in the scientific community. By the mid-1920s, scientists theorized that when light hits materials such as selenium, the more powerful photons pack enough energy to knock poorly bound electrons from their orbits. When wires are attached, the liberated electrons flow through them as electricity. Many envisioned the day when banks of selenium solar cells would power factories and light homes. However, no one could build selenium solar cells efficient enough to convert more than half a percent of the sun's energy into electricity—hardly sufficient to justify their use as a power source.

4.2 The Discovery of the Silicon Solar Cell

An accidental discovery by scientists at Bell Laboratories in 1953 revolutionized solar cell technology. Gerald Pearson and Calvin Fuller led the pioneering effort that took the silicon transistor, now the principal electronic component used in all electrical equipment, from theory to working device. Fuller had devised a way to control the introduction of impurities necessary to transform silicon from a poor to a superior conductor of electricity. He gave Pearson a piece of his intentionally contaminated silicon. Among the experiments Pearson performed using the specially treated silicon was exposing it to the sun while hooked to a device that measured electrical flow. To Pearson's surprise, he observed an electric output almost five times greater than that of the best selenium produced. He ran down the hall to tell his good friend, Daryl Chapin, who had been

trying to improve selenium to provide small amounts of intermittent power for remote locations. He said, “Don’t waste another moment on selenium!” and handed him the piece of silicon he had just tested. After a year of painstaking trial-and-error research and development, Bell Laboratories showed the public the first solar cells capable of producing useful amounts of power. Reporting the event on page 1, the *New York Times* stated that the work of Chapin, Fuller, and Pearson “may mark the beginning of a new era, leading eventually to the realization of one of mankind’s most cherished dreams—the harnessing of the almost limitless energy of the sun for the uses of civilization.”

Few inventions in the history of Bell Laboratories evoked as much media attention and public excitement as its unveiling of the silicon solar cell. Commercial success, however, failed to materialize due to the solar cells’ prohibitive costs. Desperate to find marketable products run by solar cells, manufacturers used them to power novelty items such as toys and the newly developed transistor radio. With solar cells powering nothing but playthings, one of the inventors of the solar cell, Daryl Chapin, could not hide his disappointment, wondering, “What to do with our new baby?”

4.3 The First Practical Application of Silicon Solar Cells

Unbeknown to Chapin at the time, powerful backing of the silicon solar cell was developing at the Pentagon. In 1955, the U.S. government announced its intention to launch a satellite. The prototype had silicon solar cells for its power plant. Since power lines could not be strung out to space, satellites needed a reliable, long-lasting autonomous power source. Solar cells proved the perfect answer.

The launching of the Vanguard (Fig. 8), the first satellite equipped with solar cells, demonstrated their value. Preceding satellites, run by batteries, lost power after approximately 1 week, rendering millions of dollars’ worth of equipment useless. In contrast, the solar-powered Vanguard continued to communicate with Earth for many years, allowing the completion of many valuable experiments. The success of the Vanguard’s solar power pack broke down the existing prejudice at the time toward the use of solar cells in space. As the space race between the United States and the Soviet Union intensified, both adversaries urgently needed solar cells. The demand opened a relatively large business for companies manufacturing them. More important,

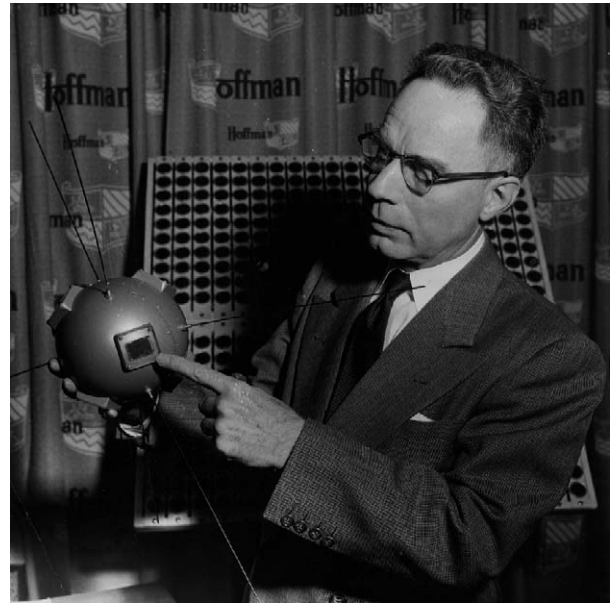


FIGURE 8 An engineer holds the Vanguard 1, the first satellite to use photovoltaics. From Perlin/Butti Solar Archives.

for the first time in the history of solar power, the sun’s energy proved indispensable to society: Without the secure, reliable electricity provided by photovoltaics, the vast majority of space applications so vital to our everyday lives would never have been realized.

4.4 The First Terrestrial Applications

Although things were looking up for solar cells in space in the 1960s and early 1970s, their astronomical price kept them distant from Earth. In 1968, Elliot Berman decided to quit his job as an industrial chemist to develop inexpensive solar cells to bring the technology from space to Earth. Berman prophetically saw that with a major decrease in price photovoltaics could play a significant role in supplying electrical power to locations on Earth where it is difficult to run power lines. After 18 months of searching for venture capital, Exxon executives invited Berman to join their laboratory in late 1969. Dismissing other solar scientists’ obsession with efficiency, Berman concentrated on lowering costs by using lower grade and therefore cheaper silicon and less expensive materials for packaging the modules. By decreasing the price from \$200 to \$20 per watt, solar cells could compete with power equipment needed to generate electricity distant from utility poles.

4.5 Terrestrial Applications

Oil companies were the first major customers of solar modules. They had both the need and the money. Oil rigs in the Gulf of Mexico were required to have warning lights and horns. Most relied on huge flashlight-type batteries to run them. The batteries required heavy maintenance and also had to be replaced approximately every 9 months. This required a large boat with an onboard crane or a helicopter. In contrast, a small skiff could transport the much lighter solar module and accompanying rechargeable battery, resulting in tremendous savings. By 1980, photovoltaics had become the standard power source for warning lights and horns not only on rigs in the Gulf of Mexico but also on those throughout the world.

Oil and gas companies also need small amounts of electricity to protect well casings and pipelines from corroding. Sending a current into the ground electrochemically destroys the offending molecules that cause this problem. However, many oil and gas fields both in the United States and in other areas of the world such as the Middle East and North Africa are far away from power lines but receive plenty of sunshine. In such cases, solar modules have proven the most cost-effective way of providing the needed current to keep pipes and casings corrosion free.

The cost of changing nonrechargeable batteries on the U.S. Coast Guard's buoys exceeded the buoys' original cost. Hence, Lloyd Lomer, at the time a lieutenant commander in the Guard, believed that switching to photovoltaics as a power source for buoys made economic sense. However, his superiors, insulated from competition, balked at the proposed change. Lomer continued his crusade, winning approval to install a test system in the most challenging of environments for sun-run devices—Ketchikan, Alaska. The success of the photovoltaic-powered buoy in Alaska proved Lomer's point. Still, his boss refused to budge. Going to higher authorities in government, Lomer eventually won approval to convert all buoys to photovoltaics. Almost every coast guard service in the world has followed suit.

To solar pioneer Elliot Berman, the railroads seemed to be another natural for photovoltaics. One of his salesmen convinced the Southern Railway (now Norfolk Southern) to try photovoltaics for powering a crossing signal at Rex, Georgia. These seemingly fragile cells did not impress veteran railroad workers as capable of powering much of anything. They therefore put in a utility-tied backup. However, a strange thing happened that turned quite

a few heads. That winter, the lines went down on several occasions due to heavy ice buildup on the wires, and the only electricity for miles around came from the solar array. As one skeptic remarked, "Rex, Georgia taught the Southern that solar worked!"

With the experiment at Rex a success, the Southern decided to put photovoltaics to work for the track circuitry, the railroad's equivalent of air-traffic control, keeping trains at a reasonable distance from one another to prevent head-on collisions or back-enders. The presence of a train changes the rate of flow of electricity running through the track. A decoder translates that change to throw signals and switches up and down the track to ensure safe passage for all trains in the vicinity. At remote spots along the line, photovoltaics provided the needed electricity.

Other railroads followed the Southern's lead. In the old days, telegraph and then telephone poles ran parallel to most tracks. Messages sent through these wires kept stations abreast of matters paramount to the safe and smooth functioning of a rail line. However, by the mid-1970s, wireless communications could do the same tasks. The poles became a liability and maintenance expense. Railroads started to dismantle poles. Whenever they found their track circuitry devices to be too far away from utility lines, they relied on photovoltaics.

The U.S. Army Signal Corps, which pioneered the use of solar power in space by equipping the solar pack on the Vanguard, was the first to bring photovoltaics down to Earth. In June 1960, the Corps sponsored the first transcontinental radio broadcast generated by the sun's energy to celebrate its 100th anniversary. The Corps, probably photovoltaics' strongest supporter in the late 1950s and the early 1960s, envisioned the solar broadcast would lead others to use photovoltaics to help provide power to run radio and telephone networks in remote locations.

Fourteen years later, GTE John Oades realized the Corps' dream by installing the first photovoltaic-run microwave repeater in the rugged mountains above Monument Valley, Utah. By minimizing the power needs of the microwave repeater, Oades could power the repeater with a small photovoltaic panel, saving all the expenses formerly attributed to them. His invention allowed people living in towns in the rugged American west, hemmed in by mountains, to enjoy the luxury of long-distance phone service that most other Americans took for granted. Prior to the solar-powered repeater, it was too expensive for phone companies to install cable or

lines, forcing people to drive hours, sometimes through blizzards on windy roads, just to make long-distance calls.

Australia had an even more daunting challenge to bring modern telecommunication services to its rural customers. Although approximately the same size as the United States, only 22 million people lived in Australia in the early 1970s when its government mandated Telecom Australia to provide every citizen, no matter how remotely situated, with the same radio, telephone, and television service enjoyed by urban customers living in Sydney or Melbourne.

Telecom Australia tried, without success, traditional stand-alone power systems such as generators, wind machines, and nonrechargeable batteries to run autonomous telephone receivers and transmitters for its rural customers. Fortunately, by 1974, it had another option—relatively cheap solar cells manufactured by Elliot Berman's firm, Solar Power Corporation. The first photovoltaic-run telephone was installed in rural Victoria, after which hundreds of others were installed. By 1976, Telecom Australia judged photovoltaics as the preferred power source for remote telephones. In fact, the solar-powered telephones proved so successful that engineers at Telecom Australia believed they were ready to develop large photovoltaic-run telecommunication networks linking towns such as Devil's Marbles, Tea Tree, and Bullocky Bone to Australia's national telephone and television service. As a result of photovoltaics, people in these and neighboring towns could dial long distance direct instead of having to call the operator and shout into the phone to be understood. Nor did they have to wait for newsreels to be flown to their local station to view news, hours, if not days, old. There were a total of 70 solar-powered microwave repeater networks by the early 1980s, the longest spanning 1500 miles. The American and Australian successes showed the world that solar power worked and benefited hundreds of thousands of people. In fact, by 1985 a consensus in the telecommunications field found photovoltaics to be the power system of choice for remote communications.

A few years earlier, Mali, a country located south of the Sahara desert, along with its neighbors, suffered from drought so devastating that tens of thousands of people and livestock died. The Malian government knew it could not save its population without the help of people like Father Vespieren, a French priest who lived in Mali for decades and had successfully run several agricultural schools. The government asked Vespieren to form a private

company to tap the vast aquifers that run underneath the Malian desert. Vespieren saw drilling as the easy part. His challenge was pumping. No power lines ran nearby and generators lay idle for lack of repairs or fuel. Then he heard about a water pump in Corsica that ran without moving parts, without fuel, and without a generating plant: It ran simply on energy from the sun. Vespieren rushed to see the pioneering installation. When he saw the photovoltaic pump, he knew that only this technology could save the people of Mali. In the late 1970s, Vespieren dedicated Mali's first photovoltaic-powered water pump with the following words: "What joy, what hope we experience when we see that sun which once dried up our pools now replenishes them with water!" By 1980, Mali, one of the poorest countries in the world, had more photovoltaic water pumps per capita than any other country due to Vespieren's efforts. The priest demonstrated that success required the best equipment, a highly skilled and well-equipped maintenance service, and financial participation by consumers. Investing their own money helped pay the bill for the capital equipment and ongoing maintenance but, more important, the people came to regard the panels and pumps as valued items that they would help care for. Most successful photovoltaic water pump projects in the developing world have followed Vespieren's example. When Vespieren initiated his photovoltaic water-pumping program, less than 10 photovoltaic pumps existed throughout the world. Today, tens of thousands of photovoltaic pumps provide water to people, livestock, and crops.

Despite the success of photovoltaic applications throughout the world in the 1970s and early 1980s, institutions responsible for rural electrification programs in developing countries did not consider installing solar electric panels to power villages distant from urban areas. Working from offices either in the West or in large cities in poorer countries, they only considered central power stations run by nuclear energy, oil, or coal. Not having lived "out in the bush," these "experts" did not realize the huge investment required to string wires from power plants to the multitudes residing in small villages tens of miles away. As a consequence, only large populated areas received electricity, leaving billions in the countryside without. The disparity in energy distribution helped cause the great migration to the cities in Africa, Asia, and Latin America. As a consequence, megalopolises such as Mexico City, Lagos, and Bombay arose with all the ensuing problems, such as crime,

diseases like AIDS, pollution, and poverty. The vast majority still living in the countryside do not have electricity. To have some modicum of lighting, they have had to rely on ad hoc solutions such as kerosene lamps. Those buying radios, tape cassettes, and televisions also must purchase batteries. It has become apparent that mimicking the Western approach to electrification has not worked in the developing world. Instead of trying to install wires and poles, which no utility in these regions can afford, bringing a 20- to 30-year supply of electricity contained in solar modules to consumers by animal or by vehicle makes more sense. Whereas it requires years to construct a centralized power plant and even more time for power lines to be installed to deliver electricity, it takes less than 1 day to install an individual photovoltaic system.

Ironically, the French Atomic Energy Commission pioneered electrifying remote homes in the outlying Tahitian Islands with photovoltaics. The Atomic Energy Commission believed that electrifying Tahiti would mollify the ill feeling in the region created by its testing of nuclear bombs in the South Pacific. The commission considered all stand-alone possibilities, including generators, wind machines, and biogas from the husks of coconut shells, before choosing photovoltaics. In 1983, 20% of the world's production of solar cells went to French Polynesia. By 1987, half of all the homes on these islands received their electricity from the sun.

Rural residents in Kenya believe that if they have to wait for electricity from the national utility, they will be old and gray by the time they receive it. To have electricity now, tens of thousands of Kenyans have bought photovoltaic units. Though relatively small compared to what an American would purchase, the typical modules in Kenya range from 12 to 25 W and are capable of charging a battery with enough electricity to run three low-wattage lights and a television for 3 hours after dark. Electric lighting provided by photovoltaics eliminates the noxious fumes and the threat of fire that people had to contend with when lighting their homes with kerosene lamps. Due to photovoltaics, children can study free from eye strain and tearing and from developing hacking coughs that they formerly experienced when studying by kerosene lamp. Their performance in school has soared with better lighting. Electric lighting also allows women to sew and weave at night products that generate cash. Newly gained economic power gives women more say in matters such as contraception. In fact, solar elec-

tricity has proven more effective in lowering fertility rates than have birth control campaigns.

Electricity provided by photovoltaics allows those living in rural areas to enjoy the amenities of urban areas without leaving their traditional homes for the cities. The government of Mongolia, for example, believed that to improve the lot of its nomadic citizens would require herding them into communities so the government could connect them to electricity generated by centralized power. The nomads, however, balked. Photovoltaics allowed the nomadic Mongolians to take part in the government's rural electrification program without giving up their traditional lifestyle. Whenever they move, the solar panels move too, with yurt and yaks and whatever else they value.

With the growing popularity of photovoltaics in the developing world, solar thievery has become common. Perhaps nothing better demonstrates the high value that people living in these countries place on photovoltaics than the drastic increase in solar module thefts during the past few years. Before the market for photovoltaics took off in the developing world, farmers would build an adobe fence around their photovoltaic panels to keep the livestock out. Now, razor wire curls around the top of these enclosures to bar solar outlaws.

As the price of solar cells has continued to decline, devices run by solar cells have seeped into the suburban and urban landscapes of the developed world. For example, when construction crews have to excavate to place underground electrical transmission lines, installing photovoltaics instead makes more economic sense. Economics has guided highway departments throughout the world to use solar cells for emergency call boxes along roads. In California alone, there are more than 30,000. The savings have been immediate. Anaheim, California, would have had to spend approximately \$11 million to connect its 11,000 call boxes to the power grid. Choosing photovoltaics instead, the city paid \$4.5 million. The city of Las Vegas found it cheaper to install photovoltaics to illuminate its new bus shelters than to have the adjacent street and sidewalk dug up to place new wires.

In the mid-1970s and early 1980s, when governments of developing countries began to fund photovoltaic projects, they copied the way electricity had been traditionally produced and delivered, favoring the construction of large fields of photovoltaic panels far away from where the electricity would be used and delivering the electricity by building miles of transmission lines. Others, however, questioned this

approach. When Charles Fritts built the first selenium photovoltaic module in the 1880s and boldly predicted that it would soon compete with Thomas Edison's generators, he envisioned "each building to have its own power plant." Since the late 1980s, most in the photovoltaic business have come around to Fritts's view. Instead of having to buy a huge amount of land—unthinkable in densely populated Europe and Japan—to place a field of panels, many began to ask why not turn each building into its own power station. Using photovoltaics as building material allows it to double as windows, roofing, skylights, facades, or any type of covering needed on a home or building, and it makes sense since the owner gets both building material and electrical generator in one package. Having the electrical production sited where the electricity will be used offers many advantages: Producing electricity on site eliminates the electricity lost in long-distance transmission, which ranges from approximately 30% in the best maintained lines to approximately 65% in rundown lines prevalent in the developing world; it reduces the danger of power lines being overloaded as demand increases by rerouting additional electricity directly to the users; and occasionally, such as on a hot August afternoon, when peak electrical demand empties the wires, excess electricity produced by photovoltaic-equipped homes and buildings can refill the lines with much needed electricity to prevent brownouts and blackouts.

4.6 The Future of Photovoltaics

Although the photovoltaics industry has experienced a phenomenal 20% annual growth rate during the past decade, it has just started to realize its potential. Although more than 1 million people in the developing world get their electricity from solar cells, more than 1 billion still have no electrical service. The continuing revolution in telecommunications is bringing a greater emphasis on the use of photovoltaics. As with electrical service, the expense of stringing telephone wires keeps most of the developing world without communication services that people in more developed countries take for granted. Photovoltaic-run satellites and cellular sites, or a combination of the two, offer the only hope to bridge the digital divide. Photovoltaics could allow everyone the freedom to dial up at or near home and, of course, hook up to the Internet.

Opportunities for photovoltaics in the developed world also continue to increase. In the United States

and Western Europe, hundreds of thousands of permanent or vacation homes are too distant for utility electric service. If people live more than 250 yards from a utility pole, paying the utility to string wires to their houses cost more than supplying their power needs with photovoltaics. Fourteen thousand Swiss Alpine chalets receive electricity from solar electricity, as do tens of thousands of others from Finland to Spain to Colorado.

Many campgrounds now prohibit RV owners from running their engines to power generators for appliances inside. The exhaust gases pollute, and the noise irritates other campers, especially at night. Photovoltaic panels mounted on the roofs of RVs provide the electricity needed without bothering others.

Restricting carbon dioxide emissions to help moderate global warming could result in the flow of large amounts of money from projects in which fossil fuels are burned to photovoltaic projects. The damage wrought by the 1997–1998 El Nino is an indication of the harsher weather expected as the earth warms. The anticipated increase in natural disasters brought on by a more disastrous future climate, as well as the growing number of people living in catastrophic-prone regions, makes early warning systems essential. The ultimate early warning device may consist of pilotless photovoltaic-powered weather surveillance airplanes, of which the *Helios* serves as a prototype. The *Helios* has flown higher than any other aircraft. Solar cells make up the entire top of the aircraft, which is merely a flying wing. Successors to the *Helios* will also have fuel cells on the underside of the wing. They will get their power from the photovoltaic panels throughout the day, extracting hydrogen and oxygen from the water discharged by the fuel cells the night before. When the sun sets, the hydrogen and oxygen will power the fuel cells, generating enough electricity at night to run the aircraft. Water discharged in the process will allow the diurnal cycle to begin the next morning. The tandem use of solar cells and fuel cells will allow the aircraft to stay aloft forever, far above the turbulence, watching for and tracking hurricanes and other potentially dangerous weather and other natural catastrophes.

Revolutionary lighting elements, called light-emitting diodes (LEDs), produce the same quality of illumination as their predecessor with only a fraction of the energy. LEDs therefore significantly reduce the amount of panels and batteries necessary for running lights, making photovoltaic systems less costly and less cumbersome. They have enabled

photovoltaics to replace gasoline generators in mobile warning signs used on roadways that alert motorists of lane closures and other temporary problems that drivers should know about. The eventual replacement of household lighting by LEDs will do the same for photovoltaics in homes.

When Bell Laboratories first unveiled the silicon solar cell, their publicist made a bold prediction: "The ability of transistors to operate on very low power gives solar cells great potential and it seems inevitable that the two Bell inventions will be closely linked in many important future developments that will influence the art of living." Already, the tandem use of transistors and solar cells for running satellites, navigation aids, microwave repeaters, televisions, radios, and cassette players in the developing world and a myriad of other devices has fulfilled the Bell prediction. It takes no great leap of imagination to expect the transistor/solar cell revolution to continue until it encompasses every electrical need from space to earth.

SEE ALSO THE FOLLOWING ARTICLES

Coal Industry, History of • Electricity Use, History of • Hydrogen, History of • Hydropower, History and Technology of • Manufactured Gas, History of • Materials for Solar Energy • Natural Gas, History of • Nuclear Power, History of • Oil Industry, History of • Solar Cells • Solar Fuels and Materials • Sun, Energy from • Wind Energy, History of • Wood Energy, History of

Further Reading

- Beattie, D. A. (ed.). (1997). "History and Overview of Solar Heat Technologies." MIT Press, Cambridge, MA.
- Butti, K., and Perlin, J. (1980). "Golden Thread: Twenty-five Hundred Years of Solar Architecture and Technology." Van Nostrand, New York.
- Perlin, J. (2002). "From Space to Earth: The Story of Solar Electricity." Harvard University Press, Cambridge, Massachusetts.

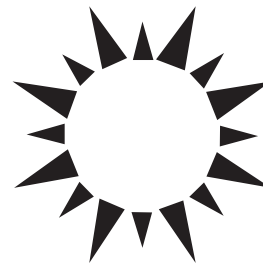
Solar Fuels and Materials

ALDO STEINFELD

ETH–Swiss Federal Institute of Technology
Zurich, Switzerland

ANTON MEIER

Paul Scherrer Institute
Villigen, Switzerland



-
1. Introduction
 2. Thermodynamics of Solar Thermochemical Conversion
 3. Solar Thermochemical Processes
 4. Outlook

Glossary

aperture Opening of a solar cavity receiver.

Carnot efficiency Maximum efficiency for converting heat from a high-temperature thermal reservoir at T_H into work in a cyclic process and rejecting heat to a low-temperature thermal reservoir at T_L , given by $1 - T_L/T_H$.

exergy efficiency (for a solar thermochemical process) The efficiency for converting solar energy into chemical energy, given by the ratio of the maximum work that may be extracted from a solar fuel to the solar energy input for producing such a fuel.

solar cavity-receiver A well-insulated enclosure, with a small opening to let in concentrated solar energy, that approaches a black-body absorber in its ability to capture solar energy.

solar concentration ratio Nondimensional ratio of the solar flux intensity (e.g., in “suns”) achieved after concentration to the incident normal beam insolation.

solar thermochemical process Any endothermic process that uses concentrated solar energy as the source of high-temperature process heat.

This article develops the underlying science of solar thermochemical processes and describes their applications for delivering clean fuels and material commodities. The thermodynamics of solar thermochemical conversion are examined, the most promising solar thermochemical processes are discussed, examples of solar chemical reactors are presented, and the latest technological developments are summarized.

1. INTRODUCTION

Using only 0.1% of the earth’s land space with solar collectors that operate with a collection efficiency of merely 20%, one could gather more than enough energy to supply the current yearly energy needs of all inhabitants of the planet ($\sim 1.2 \times 10^{14}$ kWh). Furthermore, the solar energy reserve is essentially unlimited, and its use is ecologically benign. However, solar radiation is dilute (only ~ 1 kW/m²), intermittent (available only during daytime and under clear-sky conditions), and unequally distributed (mostly by the equator). These drawbacks can be overcome by converting solar energy into chemical energy carriers—solar fuels—that can be stored long term and transported long range, from the sunny and desert regions to the industrialized and populated centers of the earth, where much of the energy is needed. Solar fuels can be burned to generate heat, further processed into electrical or mechanical work, or used directly to generate electricity in fuel cells and batteries to meet customers’ energy demands. Solar process heat can also assist in the processing and recycling of energy-intensive materials, thereby avoiding greenhouse gas emissions and other pollutants derived from the combustion of fossil fuels for heat and electricity generation.

There are basically three pathways for making solar fuels from solar energy:

- *Solar electrochemical path*: solar-made electricity, from photovoltaic or solar thermal systems, followed by an electrolytic process
- *Solar photochemical path*: direct use of the photon energy
- *Solar thermochemical path*: solar-made heat followed by a thermochemical process

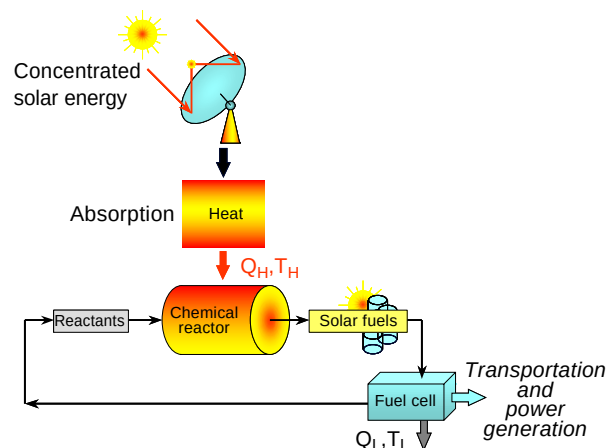


FIGURE 1 Schematic of solar energy conversion into solar fuels. Concentrated solar radiation is used as the energy source for high-temperature process heat to drive endothermic chemical reactions toward the production of storable and transportable fuels.

Combinations of these three pathways are possible, but the thermochemical path offers some intriguing thermodynamic advantages. It is this approach that is the subject of this article, whereas the other two paths are beyond the scope of the article.

Figure 1 illustrates the basic idea of the solar thermochemical path. By concentrating the diluted sunlight over a small area with help of parabolic mirrors and then capturing the radiant energy with help of suitable receivers and reactors, we can obtain heat at high temperatures for carrying out an endothermic chemical transformation and producing storable and transportable fuels. These solar fuels ultimately store solar energy within their chemical bonds. Regardless of the fuel, the higher the temperature of the reaction process, the higher the efficiency of energy conversion. However, higher temperatures also lead to greater losses by reradiation from the solar receiver. The thermodynamic implications of such phenomena are examined in the following section.

2. THERMODYNAMICS OF SOLAR THERMOCHEMICAL CONVERSION

The state-of-the-art technology for large-scale collection and concentration of solar energy is based on three main optical configurations using parabolic-shaped mirrors. These are the trough, tower, and dish systems shown schematically in Fig. 2. Trough

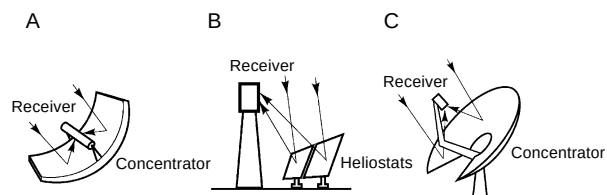


FIGURE 2 Schematic of the three main optical configurations for large-scale collection and concentration of solar energy: (A) the trough system, (B) the tower system, and (C) the dish system.

systems use linear, two-dimensional, parabolic mirrors to focus sunlight onto a solar tubular receiver positioned along their focal line. Tower systems use a field of heliostats (two-axis tracking parabolic mirrors) that focus the sunrays onto a solar receiver mounted on top of a centrally located tower. Dish systems use paraboloidal mirrors to focus sunlight on a solar receiver positioned at their focus. A recently developed Cassegrain optical configuration for the tower system makes use of a hyperboloidal reflector at the top of the tower to redirect sunlight to a receiver located on the ground level. Solar furnaces are concentrating facilities in which high-flux solar intensities are usually obtained at a fixed location inside a housed laboratory. The traditional design consists of using a sun-tracking, flat heliostat on-axis with a stationary primary paraboloidal concentrator; off-axis configurations have also been designed.

The capability of these collection systems to concentrate solar energy is often expressed in terms of their mean flux concentration ratio $\tilde{C}$ over a targeted area A at the focal plane, normalized with respect to the incident normal beam insolation I as follows:

$$\tilde{C} = \frac{Q_{\text{solar}}}{I \cdot A}, \quad (1)$$

where Q_{solar} is the solar power input into the target. $\tilde{C}$ is often expressed in units of “suns” when normalized to 1 kW/m^2 . The solar flux concentration ratio typically obtained at the focal plane varies between 30 and 100 suns for trough systems, between 500 and 5,000 suns for tower systems, and between 1,000 and 10,000 suns for dish systems. Higher concentration ratios imply lower heat losses from smaller receivers and, consequently, higher attainable temperatures at the receiver. To some extent, the flux concentration can be further augmented with the help of a nonimaging secondary concentrator (e.g., a compound parabolic concentrator [CPC]) that is positioned in tandem with the primary concentrating system. The aforementioned

solar concentrating systems have proved to be technically feasible in large-scale experimental demonstrations aimed mainly at the production of solar thermal electricity in which a working fluid (typically air, water, helium, sodium, or molten salt) is solar heated and further used in traditional Rankine, Brayton, and Stirling cycles. Solar thermochemical applications, although not developed to the extent that solar thermal electricity generation has been, employ the same solar concentrating technology.

Solar chemical reactors for highly concentrated solar systems usually feature the use of a cavity receiver-type configuration, that is, a well-insulated enclosure with a small opening (the aperture) to let in concentrated solar radiation. Because of multiple internal reflections, the fraction of the incoming energy absorbed by the cavity greatly exceeds the simple surface absorptance of the inner walls. The larger the ratio of the cavity's characteristic length to the aperture diameter, the closer the cavity receiver approaches a black-body absorber. Smaller apertures also serve to reduce reradiation losses. However, they intercept a reduced fraction of the sunlight reflected from the concentrators. Consequently, the optimum aperture size is a compromise between maximizing radiation capture and minimizing radiation losses.

The solar energy absorption efficiency of a solar reactor, $\eta_{\text{absorption}}$, is defined as the net rate at which energy is being absorbed divided by the solar power coming from the concentrator. At temperatures above approximately 1000 K, the net power absorbed is diminished mostly by radiant losses through the aperture. For a perfectly insulated black-body cavity receiver, it is given by

$$\eta_{\text{absorption}} = 1 - \left(\frac{\sigma T^4}{I\tilde{C}} \right), \quad (2)$$

where T is the nominal cavity-receiver temperature and σ is the Stefan–Boltzmann constant. The absorbed concentrated solar radiation drives an endothermic chemical reaction. The measure of how well solar energy is converted into chemical energy for a given process is the exergy efficiency, defined as

$$\eta_{\text{exergy}} = \frac{-\dot{n}\Delta G|_{298\text{ K}}}{Q_{\text{solar}}}, \quad (3)$$

where Q_{solar} is the solar power input, $\dot{n}$ is the products' molar flow rate, and ΔG is the maximum possible amount of work (Gibbs free energy change) that may be extracted from the products as they are transformed back to reactants at 298 K. The second law of thermodynamics is now applied to calculate

the theoretical maximum exergy efficiency $\eta_{\text{exergy,ideal}}$. Because the conversion of solar process heat to chemical energy is limited by both the solar absorption efficiency and the Carnot efficiency, the maximum ideal exergy efficiency is given by

$$\eta_{\text{exergy,ideal}} = \eta_{\text{absorption}} \cdot \eta_{\text{Carnot}} = \left[1 - \left(\frac{\sigma T_H^4}{I\tilde{C}} \right) \right] \times \left[1 - \left(\frac{T_L}{T_H} \right) \right], \quad (4)$$

where T_H and T_L are the upper and lower operating temperatures of the equivalent Carnot heat engine. $\eta_{\text{exergy,ideal}}$ is plotted in Fig. 3 as a function of T_H for $T_L = 298$ K and for various solar flux concentration ratios. Because of the Carnot limit, one should try to operate thermochemical processes at the highest upper temperature possible; however, from a heat transfer perspective, the higher the temperature, the higher the reradiation losses. The highest temperature an ideal solar cavity receiver is capable of achieving, defined as the stagnation temperature $T_{\text{stagnation}}$, is calculated by setting $\eta_{\text{exergy,ideal}}$ equal to zero, yielding

$$T_{\text{stagnation}} = \left(\frac{I\tilde{C}}{\sigma} \right)^{0.25}. \quad (5)$$

At this temperature, energy is being reradiated as fast as it is absorbed. Stagnation temperatures exceeding 3000 K are attainable with solar concentration ratios above 5000 suns. However, an energy-efficient

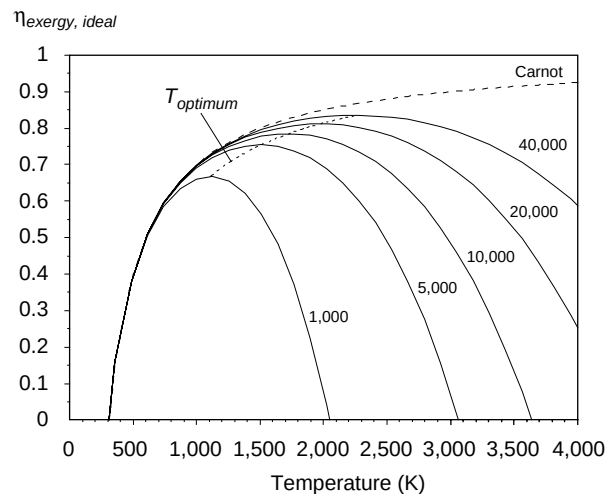


FIGURE 3 Variation of the ideal exergy efficiency $\eta_{\text{exergy,ideal}}$ as a function of the operating temperature T_H for a black-body cavity receiver converting concentrated solar energy into chemical energy. The mean solar flux concentration is the parameter: 1,000 ... 40,000. Also plotted are the Carnot efficiency and the locus of the optimum cavity temperature T_{optimum} .

process must run at temperatures that are substantially below $T_{stagnation}$. There is an optimum temperature $T_{optimum}$ for maximum efficiency obtained by setting

$$\frac{\partial \eta_{exergy,ideal}}{\partial T} = 0. \quad (6)$$

Assuming a uniform power flux distribution, this relation yields the following implicit equation for $T_{optimum}$:

$$T_{optimum}^5 - \frac{3}{4} T_L T_{optimum}^4 - \frac{T_L I \tilde{C}}{4\sigma} = 0. \quad (7)$$

The locus of $T_{optimum}$ is shown in Fig. 3. For example, for a solar concentration ratio of 5000, the optimum temperature of a solar receiver is 1500 K, giving a maximum theoretical efficiency of 75%, that is, the portion of solar energy that could in principle be converted into the chemical energy of fuels. In practice, when considering convection and conduction losses in addition to reradiation losses, the efficiency will peak at a somewhat lower temperature.

The process modeling described in the following paragraphs establishes a base for evaluating and comparing different solar thermochemical processes in terms of their ideal exergy efficiencies. Figure 4 depicts a flow diagram of an ideal cyclic process that leads to a fuel, which uses a solar reactor, a quenching device, and a fuel cell. The complete

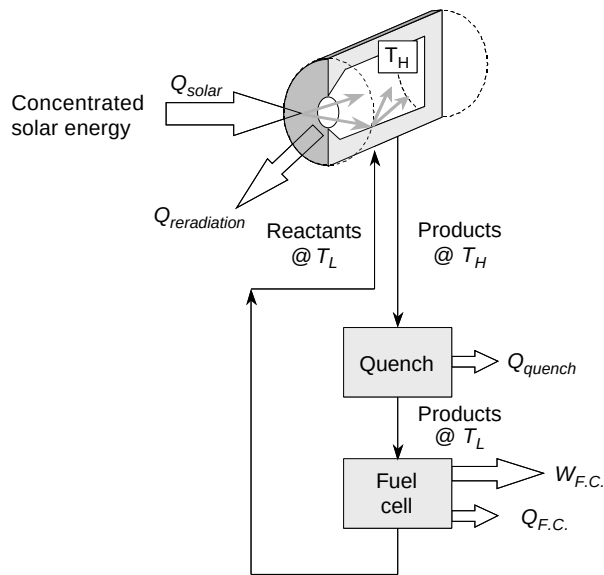


FIGURE 4 Model flow diagram of an ideal cyclic process for calculating the maximum exergy efficiency of a solar thermochemical process.

process is carried out at constant pressure. In practice, pressure drops will occur throughout the system. However, if one assumes frictionless operating conditions, no pumping work is required. The reactants may be preheated in an adiabatic heat exchanger, where some portion of the sensible and latent heat of the products is transferred to the reactants; for simplicity, a heat exchanger has been omitted. The reactor is assumed to be a perfect black-body cavity receiver. The reactants enter the solar reactor at T_L and are further heated to the reaction temperature T_H . Chemical equilibrium is assumed inside the reactor. The net power absorbed in the solar reactor should match the enthalpy change per unit time of the reaction,

$$Q_{reactor,net} = \dot{n}\Delta H |_{\text{Reactants}@T_L \rightarrow \text{Products}@T_H}. \quad (8)$$

Irreversibility in the solar reactor arises from the nonreversible chemical transformation and reradiation losses $Q_{reradiation}$ to the surroundings at T_L . It is found that

$$Irr_{reactor} = \left(\frac{-Q_{solar}}{T_H} \right) + \left(\frac{Q_{reradiation}}{T_L} \right) + \left(\dot{n}\Delta S |_{\text{Reactants}@T_L \rightarrow \text{Products}@T_H} \right). \quad (9)$$

Products exit the solar reactor at T_H and are cooled rapidly to T_L . The amount of power lost during quenching is

$$Q_{quench} = -\dot{n}\Delta H |_{\text{Products}@T_H \rightarrow \text{Products}@T_L} \quad (10)$$

The irreversibility associated with quenching is

$$Irr_{quench} = \left(\frac{Q_{quench}}{T_L} \right) + \left(\dot{n}\Delta S |_{\text{Products}@T_H \rightarrow \text{Products}@T_L} \right). \quad (11)$$

Introducing a reversible ideal fuel cell, in which the products recombine to form the original reactants and thereby generate electrical power, closes the cycle. The maximum amount of work that the products leaving the reactor could produce if they combined at T_L and a total pressure of 1 bar is given by

$$W_{F.C.} = -\dot{n}\Delta G |_{\text{Products}@T_L \rightarrow \text{Reactants}@T_L}. \quad (12)$$

This work value is also known as the exergy of the products at ambient temperature. The fuel cell operates isothermally; the amount of heat rejected to the surroundings is

$$Q_{F.C.} = -T_L \cdot \dot{n}\Delta S |_{\text{Products}@T_L \rightarrow \text{Reactants}@T_L}. \quad (13)$$

The exergy system efficiency of the closed cycle is then calculated using Eq. (3) as

$$\eta_{\text{exergy}} = \frac{W_{\text{F.C.}}}{Q_{\text{solar}}} \quad (14)$$

This thermodynamic analysis can be verified by performing an energy balance and by evaluating the maximum achievable efficiency (Carnot efficiency) from the total available work and from the total power input. The energy balance should satisfy

$$W_{\text{F.C.}} = Q_{\text{solar}} - (Q_{\text{reradiation}} + Q_{\text{quench}} + Q_{\text{F.C.}}). \quad (15)$$

The available work is calculated as the sum of the fuel cell work plus the lost work due to the irreversibility associated with the solar reactor and with quenching. Thus,

$$\begin{aligned} \eta_{\text{max}} &= \frac{W_{\text{F.C.}} + T_L \cdot (Irr_{\text{reactor}} + Irr_{\text{quench}})}{Q_{\text{solar}}} \\ &= 1 - \frac{T_L}{T_H} = \eta_{\text{Carnot}}. \end{aligned} \quad (16)$$

Any solar thermochemical processes can be thought of in this manner, and their exergy efficiencies can be compared as one criterion for judging their relative industrial potentials. The higher the exergy efficiency, the lower the required solar collection area for producing a given amount of solar fuel and, consequently, the lower the costs incurred for the solar concentrating system, which usually correspond to half of the total investments for the entire solar chemical plant. Thus, high-exergy efficiency implies favorable competitiveness.

3. SOLAR THERMOCHEMICAL PROCESSES

3.1 Solar Production of Fuels

3.1.1 Solar Hydrogen

Five thermochemical routes for solar hydrogen production are depicted in Fig. 5. Indicated is the chemical source of H_2 : water for the solar thermolysis and solar thermochemical cycles, fossil fuels for the solar cracking, and a combination of fossil fuels and H_2O for the solar reforming and solar gasification. All of these routes make use of concentrated solar radiation as the energy source of high-temperature process heat.

3.1.1.1 H_2 from H_2O Although the single-step thermal dissociation of water, known as water thermolysis, is conceptually simple, it has been

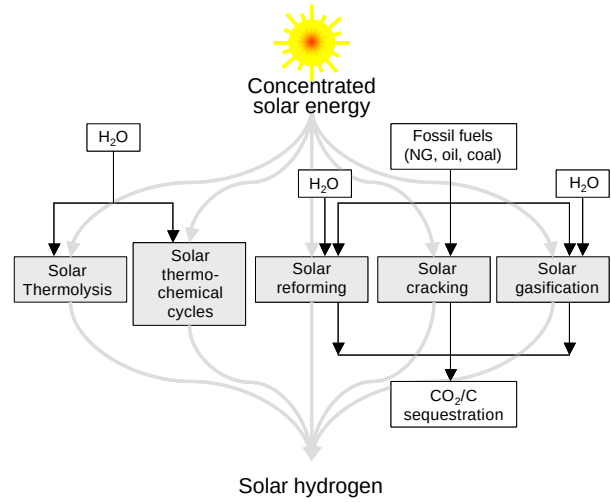
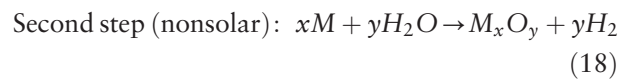
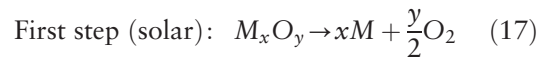


FIGURE 5 Five thermochemical routes for the production of solar hydrogen.

impeded by the need for a high-temperature heat source at above 2500 K to achieve a reasonable degree of dissociation and by the need for an effective technique to separate H_2 and O_2 to avoid recombination or ending up with an explosive mixture. Among the ideas proposed for separating H_2 from the products are effusion and electrolytic separation. Water-splitting thermochemical cycles bypass the H_2/O_2 separation problem and further allow operating at relatively moderate upper temperatures. Previous studies performed on H_2O -splitting thermochemical cycles were characterized mostly by the use of process heat at temperatures below approximately 1300 K, available from nuclear and other thermal sources. These cycles required multiple (more than two) steps and were suffering from inherent inefficiencies associated with heat transfer and product separation at each step. During recent years, significant progress has been made in the development of optical systems for large-scale collection and concentration of solar energy capable of achieving solar concentration ratios of 5000 suns or higher. Such high-radiation fluxes allow the conversion of solar energy to thermal reservoirs at 2000 K and above that are needed for the two-step thermochemical cycles using metal oxide redox reactions:



Here, M denotes a metal and $M_x\text{O}_y$ denotes the corresponding metal oxide. The cycle is represented

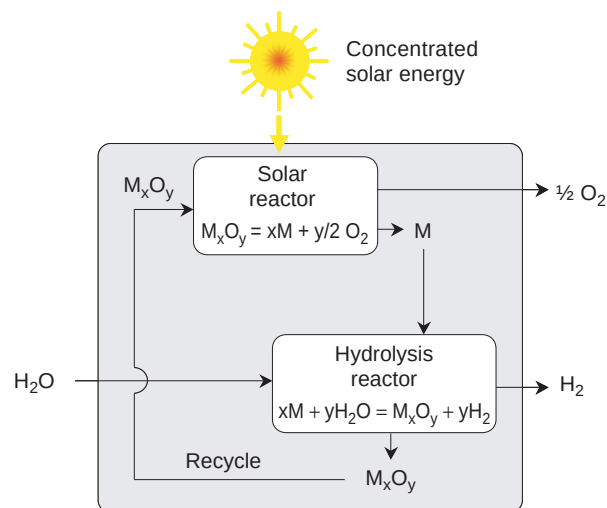


FIGURE 6 Representation of a two-step water-splitting thermochemical cycle using metal oxide redox reactions.

in Fig. 6. The first endothermic step is the solar thermal dissociation of the metal oxide to the metal or the lower valence metal oxide. The second, nonsolar exothermic step is the hydrolysis of the metal to form H_2 and the corresponding metal oxide. The net reaction is $H_2O = H_2 + 0.5O_2$, but because H_2 and O_2 are formed in different steps, the need for high-temperature gas separation is thereby eliminated. The most favorable candidate metal oxide redox pairs for this cycle are ZnO/Zn and Fe_3O_4/FeO . In both cases, the minimum temperature requirement for the solar step is 2300 K, and the products need to be either quenched or separated at high temperatures to prevent their recombination. Quenching introduces irreversibility and may be a factor of complexity in large-scale use.

Figure 7 shows a solar chemical reactor concept for performing the thermal dissociation of ZnO that features a windowed rotating cavity receiver lined with ZnO particles that are held by centrifugal force. With this arrangement, ZnO is directly exposed to high-flux solar irradiation and serves the functions of radiant absorber, thermal insulator, and chemical reactant simultaneously. The shell of the cavity is made from conventional steel materials and is lined with the same material as are the reactants themselves. This aspect of the design eliminates the need for using expensive and difficult-to-fabricate ceramic insulating materials for ultra-high temperatures. It also offers excellent resistance to thermal shocks that are intrinsic in short start-up solar applications. The direct absorption of concentrated solar energy by directly irradiated reactants

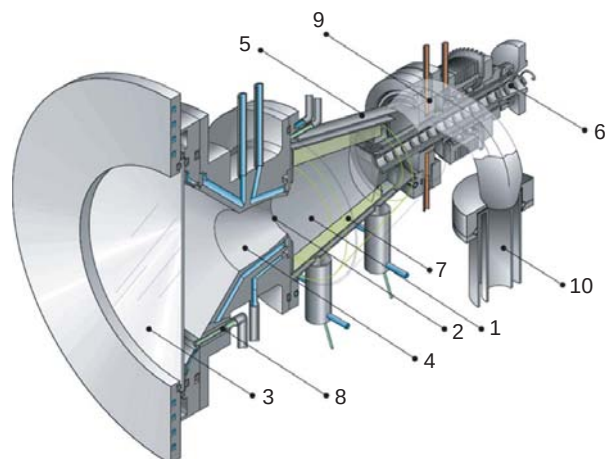


FIGURE 7 Schematic of a “rotating-cavity” solar chemical reactor for the thermal dissociation of ZnO to Zn and O_2 at 2300 K. It consists of a rotating conical cavity receiver (1) that contains an aperture (2) for access of concentrated solar radiation through a quartz window (3). The solar flux concentration is further augmented by incorporating a CPC (4) in front of the aperture. Both the window mount and the CPC are water cooled and integrated into a concentric (nonrotating) conical shell (5). ZnO particles are fed continuously by means of a screw powder feeder located at the rear of the reactor (6). The centripetal acceleration forces the ZnO powder to the wall, where it forms a thick layer of ZnO (7) that insulates and reduces the thermal load on the inner cavity walls. A purge gas flow enters the cavity receiver tangentially at the front (8) and keeps the window cool and clear of particles or condensable gases. The gaseous products Zn and O_2 exit continuously via an outlet port (9) to a quench device (10). Source: Paul Scherrer Institute, Villigen, Switzerland.

provides efficient radiation heat transfer to the reaction site where the energy is needed, bypassing the limitations imposed by indirect heat transport via heat exchangers.

We illustrate the application of the Second-Law (exergy) analysis, Eqs. (1) to (16), for calculating the theoretical maximum exergy efficiency of the two-step water-splitting solar thermochemical cycle based on the ZnO/Zn redox reaction. Figure 8 depicts the ideal cyclic process. The solar reactor is assumed to be a cavity-receiver operating at 2300 K. The reactant $ZnO(s)$ enters the solar reactor at 298 K and is further heated to the reaction temperature at 2300 K. The molar feed rate of ZnO to the solar reactor, $\dot{n}$, is set to 1 mol/s and is equal to that of H_2O fed to the hydrolyser reactor. After quenching, the products separate naturally (without expending work) into gaseous O_2 and condensed phase Zn . Zinc is sent to the hydrolyser reactor to react exothermally with water and form hydrogen, according to Eq. (18). The heat liberated is assumed

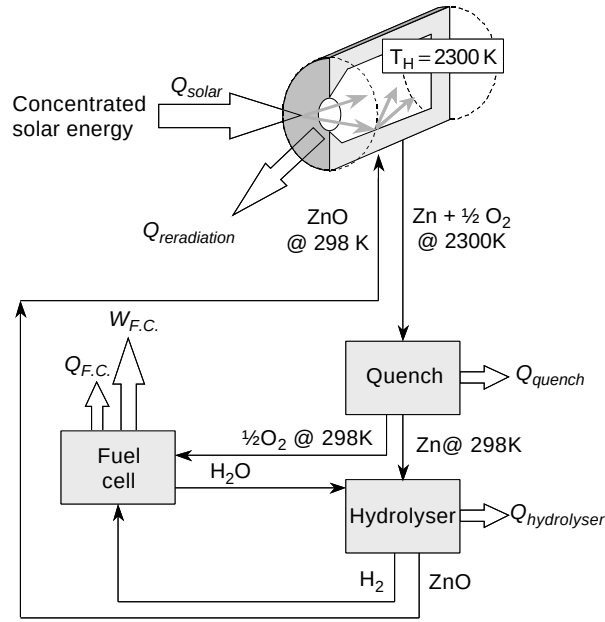


FIGURE 8 Model flow diagram of the two-step water-splitting solar thermochemical cycle used for the exergy analysis.

lost to the surroundings, as given by

$$Q_{\text{hydrolyser}} = -\dot{n}\Delta H|_{\text{Zn}+\text{H}_2\text{O}@298\text{ K}\rightarrow\text{ZnO}+\text{H}_2@298\text{ K}} \quad (19)$$

Thus, the irreversibility associated with the hydrolyser is

$$\begin{aligned} Irr_{\text{hydrolyser}} = & \left(\frac{Q_{\text{hydrolyser}}}{298\text{ K}} \right) \\ & + \left(\dot{n}\Delta S|_{\text{Zn}+\text{H}_2\text{O}@298\text{ K}\rightarrow\text{ZnO}+\text{H}_2@298\text{ K}} \right). \end{aligned} \quad (20)$$

Introducing an H_2/O_2 fuel cell, in which the products recombine to form H_2O and thereby generate electrical power, closes the cycle. $W_{\text{F.C.}}$ and $Q_{\text{F.C.}}$ are the work output and the heat rejected, respectively, given by

$$W_{\text{F.C.}} = -\eta_{\text{F.C.}} \dot{n}\Delta G|_{\text{H}_2+0.5\text{O}_2@298\text{ K}\rightarrow\text{H}_2\text{O}@298\text{ K}} \quad (21)$$

$$Q_{\text{F.C.}} = -T_L \dot{n}\Delta S|_{\text{H}_2+0.5\text{O}_2@298\text{ K}\rightarrow\text{H}_2\text{O}@298\text{ K}}, \quad (22)$$

where $\eta_{\text{F.C.}}$ is the efficiency of the H_2/O_2 fuel cell, assumed to be 100% when based on the maximum work output of H_2 (also known as its exergy). Finally, the exergy efficiency is calculated using Eq. (14). Table I gives the results of the calculation for two values of the solar concentration ratio: 5,000 and 10,000.

3.1.1.2 H_2 from Fossil Fuels Three solar thermochemical processes for H_2 production using fossil

TABLE I

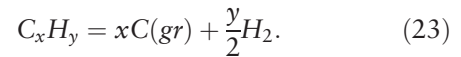
Exergy Analysis of the Two-Step Water-Splitting Solar Thermochemical Cycle Using the Process Modeling Depicted in Fig. 8

$\bar{C}$	5,000	10,000
Q_{solar}	815 kW	662 kW
$Q_{\text{reradiation}}$	258 kW	105 kW
$Q_{\text{reactor,net}}$	557 kW	
Irr_{reactor}	0.81 kW · K ⁻¹	
Q_{quench}	209 kW	
Irr_{quench}	0.52 kW · K ⁻¹	
$Q_{\text{hydrolyser}}$	64 kW	
$Irr_{\text{hydrolyser}}$	0.27 kW · K ⁻¹	
$Q_{\text{F.C.}}$	49 kW	
$W_{\text{F.C.}}$	237 kW	
$\eta_{\text{absorption}}$	68%	84%
η_{exergy} (without heat recovery)	29%	36%
η_{exergy} (with heat recovery)	69%	84%

Note. The ZnO and H_2O molar rates are set to 1 mol/s.

fuels as the chemical source are considered: cracking, reforming, and gasification. These routes are shown schematically in Fig. 9.

The solar cracking route refers to the thermal decomposition of natural gas (NG), oil, and other hydrocarbons and can be represented by the simplified net reaction:



Other compounds may also be formed, depending on the reaction kinetics and on the presence of impurities in the raw materials. The thermal decomposition yields a carbon-rich condensed phase and a hydrogen-rich gas phase. The carbonaceous solid product can either be sequestered without CO_2 release or be used as a material commodity under less severe CO_2 restraints. It can also be applied as a reducing agent in metallurgical processes. The hydrogen-rich gas mixture can be further processed to high-purity hydrogen that is not contaminated with oxides of carbon and, thus, can be used in proton exchange membrane (PEM) fuel cells without inhibiting platinum-made electrodes. From the point of view of carbon sequestration, it is easier to separate, handle, transport, and store solid carbon than gaseous CO_2 . Furthermore, while the steam-reforming/gasification method requires additional steps for shifting CO and for separating CO_2 , the thermal cracking accomplishes the removal and separation of carbon in a single step. In contrast,

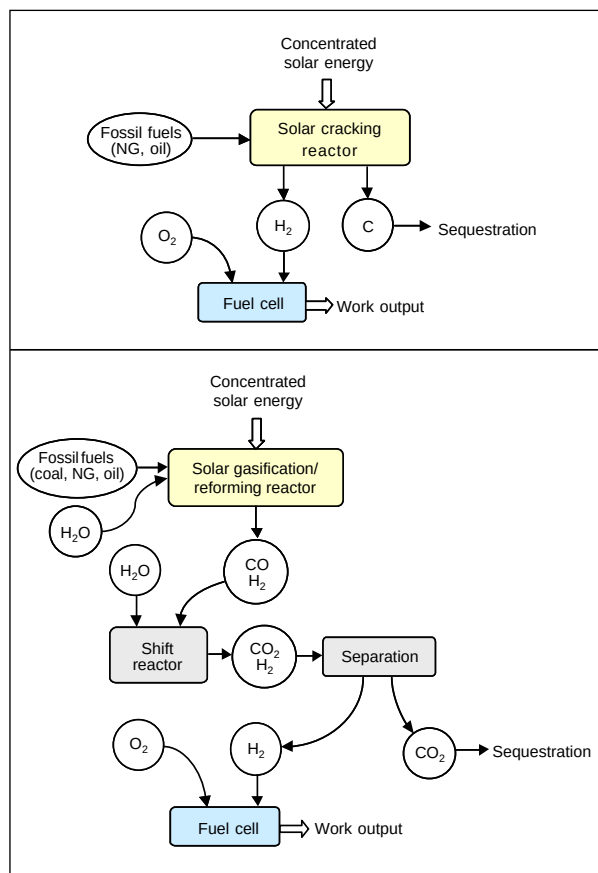


FIGURE 9 Schematic of solar thermochemical routes for H_2 production using fossil fuels and H_2O as the chemical source: solar cracking (upper box) and solar reforming and gasification (lower box).

the major drawback of the thermal decomposition method is the energy loss associated with the sequestration of carbon. Thus, the solar cracking may be the preferred option for NG and other hydrocarbons with high H_2 /carbon ratios. The exergy efficiency for these routes is defined as the ratio of the work output by the fuel cell to the total thermal energy input by both the solar energy and the heating value of the reactants:

$$\eta_{\text{exergy}} = \frac{W_{\text{F.C.}}}{Q_{\text{solar}} + HHV_{\text{reactant}}}. \quad (24)$$

where HHV_{reactant} is the high heating value of the fossil fuel being processed (e.g., $\sim 890 \text{ kJ} \cdot \text{mol}^{-1}$ for NG). Assuming a 65% efficient H_2/O_2 fuel cell, $\eta_{\text{exergy}} = 30\%$. This route offers zero CO_2 emissions as a result of carbon sequestration. However, the energy penalty for completely avoiding CO_2 reaches 30% of the electrical output vis-à-vis the direct use of NG for fueling a 55% efficient combined Brayton–

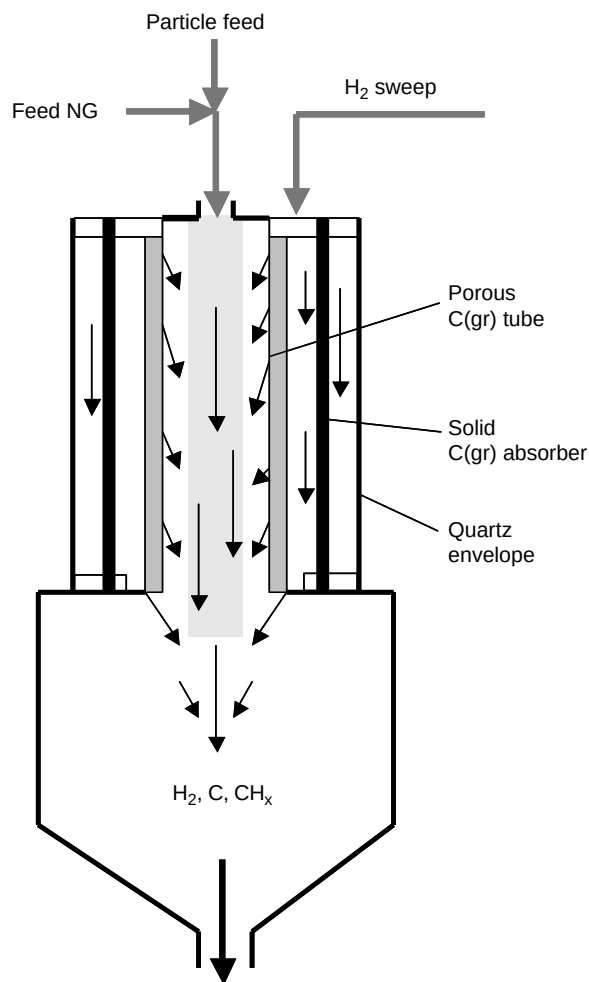
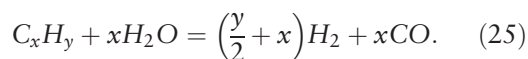


FIGURE 10 Scheme of an “aerosol” solar reactor for cracking NG. Source: National Renewable Energy Laboratory, Golden, CO.

Rankine cycle. Figure 10 shows an “aerosol” solar reactor concept for cracking NG that features two concentric graphite tubular reactors, with the outer solid tube serving as the solar absorber and the inner porous tube containing a flow of NG laden with carbon nanoparticles that serve the functions of radiant absorbers.

The steam reforming of NG, oil, and other hydrocarbons, as well as the steam gasification of coal and other carbonaceous materials, can be represented by the simplified net reaction:



Other compounds may also be formed, especially with coal, but some impurities contained in the raw materials are cleaned out prior to the decarbonization process. The principal product is high-quality synthesis gas (syngas), the building block for a wide

variety of synthetic fuels, including Fischer–Tropsch-type chemicals, hydrogen, ammonia, and methanol. Its quality is determined mainly by the H_2/CO and CO_2/CO molar ratios. For example, the solar steam gasification of anthracite coal at above 1500 K yields syngas with an H_2/CO molar ratio of 1.2 and a CO_2/CO molar ratio of 0.01. The CO content in the syngas can be shifted to H_2 via the catalytic water–gas shift reaction ($CO + H_2O = H_2 + CO_2$), and the product CO_2 can be separated from H_2 using, for example, the pressure swing adsorption technique. The exergy efficiency for this route is defined by Eq. (24). Assuming a 65% efficient H_2/O_2 fuel cell and $HHV_{reactant}$ of $35,700 \text{ kJ kg}^{-1}$ for anthracite coal, $\eta_{exergy} = 46\%$. This route offers a net gain in the electrical output by a factor varying in the range of 1.7 to 1.8 (depending on the coal type) vis-à-vis the direct use of coal for fueling a 35% efficient Rankine cycle. If CO_2 is not sequestered, specific CO_2 emissions amount to 0.53 to 0.56 kg CO_2/kWh_e , approximately half as much as the specific CO_2 emissions discharged by conventional coal-fired power plants.

Some of these processes are practiced at an industrial scale, with the process heat supplied by burning a significant portion of the feedstock. Internal combustion results in the contamination of the gaseous products, whereas external combustion results in a lower thermal efficiency due to the irreversibility associated with indirect heat transfer.

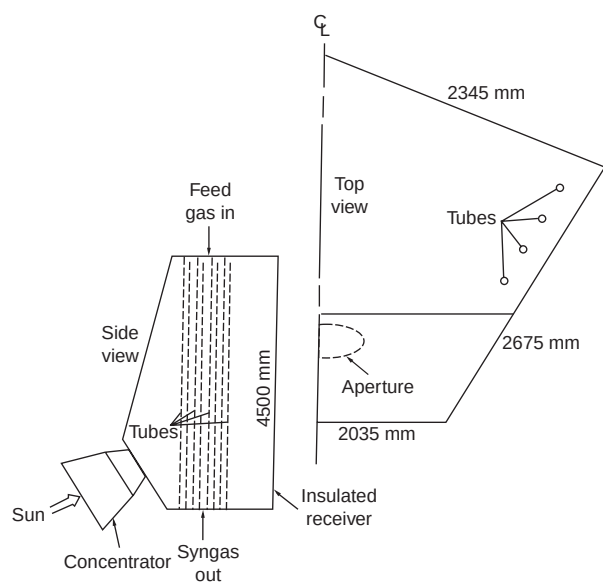


FIGURE 11 Scheme of an indirect-irradiated solar reforming chemical reactor. Source: Weizmann Institute of Science, Rehovot, Israel.

Alternatively, using solar energy for process heat offers a threefold advantage: (1) the discharge of pollutants is avoided, (2) the gaseous products are not contaminated, and (3) the calorific value of the fuel is upgraded by adding solar energy in an amount equal to the ΔH of the reaction.

The solar reforming of NG has been studied extensively in solar furnaces with small-scale solar reactor prototypes using Rh-based catalyst. Recently, this solar process has been scaled up to power levels of 300 to 500 kW in a solar tower facility using two solar reactor concepts: indirect irradiation and direct irradiation.

Figure 11 shows a schematic of the indirect-irradiated solar reforming reactor. It consists of a pentagonal cavity receiver, insulated with ceramic fibers, containing a set of Inconel tubes arranged vertically in two banks parallel to the front-side walls. The tubes are filled with a packed bed of catalyst, usually 2% Rh on Al_2O_3 support. The tubular reactors are externally exposed to the concentrated solar radiation entering through the cavity aperture and transfer the absorbed heat to the chemical reactants. The aperture is windowless and tilted from the horizontal axis for an optimal view to the heliostat field. A matching CPC is implemented at the aperture for capturing radiation spillage, augmenting the average solar flux concentration, and providing uniform heating of the tubes. Indirect-irradiated reactors such as this one have the advantage of eliminating the need for a transparent window. The disadvantages are linked to the limitations imposed by the materials of construction of the reactor walls: limitations in the maximum operating temperature, thermal conductivity, radiant absorptance, inertness, resistance to thermal shocks, and suitability for transient operation.

Figure 12 shows a schematic of the direct-irradiated solar reforming reactor, also referred to

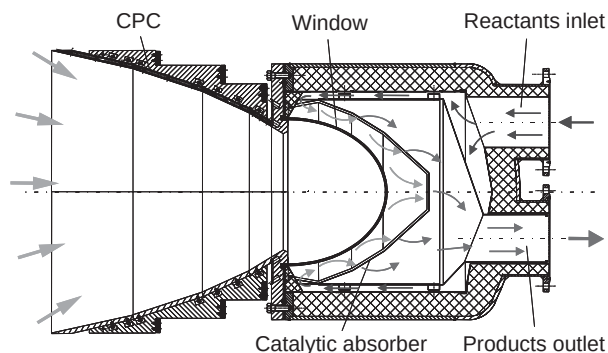


FIGURE 12 Scheme of a direct-irradiated solar reforming chemical reactor. Source: Deutsches Zentrum für Luft- und Raumfahrt, Germany.

as the volumetric reactor. The main component is the porous ceramic absorber, coated with Rh catalyst, that is directly exposed to the concentrated solar radiation. A concave quartz window, mounted at the aperture, minimizes reflection losses and permits operation at high pressures. Similar to the indirect-irradiated reactor, a CPC is implemented at the aperture. Direct-irradiated reactors such as this one have the advantage of providing efficient radiation heat transfer directly to the reaction. Furthermore, under proper conditions, direct irradiation may photochemically enhance the kinetics of the reaction. The major drawback, when working with reducing or inert atmospheres, is the requirement of a transparent window, which is a critical and troublesome component under high-pressure and severe gas environment.

The solar dry reforming of CH_4 (with CO_2) can be applied in a closed-loop chemical heat pipe (Fig. 13). The product of this reversible reaction is syngas that can be stored at ambient temperatures and transported to the site where the energy is needed. By the reverse exothermic reaction, stored solar energy is released in the form of high-temperature heat, which can be used, for example, for generating electricity via a Rankine cycle. The products of this reverse reaction are again recycled to the solar reactor, where the process is repeated. The dissociation/synthesis of NH_3 can also be applied in a chemical heat pipe for storage and transportation of solar energy.

An optional source of H_2 is H_2S , a highly toxic industrial product recovered in large quantities in the sweetening of NG and in the removal of organically

bound sulfur from petroleum and coal. Current industrial practice uses the Claus process to recover the sulfur from H_2S , but the process wastes H_2 to produce low-grade process heat. Alternatively, H_2S can be decomposed to H_2 and S_2 at 1800 K and 0.03 to 0.5 bar. The product gas mixture is quenched, and condensed S_2 is separated from H_2 . In contrast to H_2O thermolysis, the quench is relatively easy and the reverse reaction between the products seems to be unimportant at temperatures as high as 1500 K.

3.1.2 Solar Metals

Metals are attractive candidates for storage and transport of solar energy. They may be used to generate either high-temperature heat via combustion or electricity via fuel cells and batteries. Metals can also be used to produce hydrogen from H_2O via hydrolysis, as shown in Eq. (18). The chemical products from any of these power/ H_2 -generating processes are metal oxides, which in turn need to be recycled. The conventional extraction of metals from their oxides by carbothermic and electrolytic processes is characterized by its high energy consumption and its concomitant environmental pollution. The extractive metallurgical industry discharges vast amounts of greenhouse gases and other pollutants to the environment, derived mainly from the combustion of fossil fuels for heat and electricity generation. These emissions can be reduced substantially, or even eliminated completely, by using concentrated solar energy as the source of high-temperature process heat.

The thermal dissociation and electrothermal reduction of metal oxides proceeds without reducing agents, while the carbothermal reduction of metal oxides uses solid carbon $\text{C}(\text{gr})$ or hydrocarbons (e.g., CH_4) as reducing agents. The corresponding overall chemical reactions may be represented as follows:

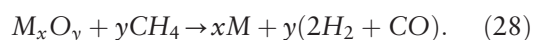
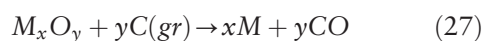
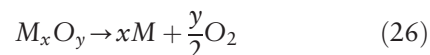


Table II lists the approximate temperatures at which the standard $\Delta G_{\text{rxn}}^\circ$ for Reactions (26), (27), and (28) equals zero for various metal oxides of interest.

Except for the thermal dissociation of ZnO , the required temperature for effecting Reaction (26) exceeds 3000 K. At temperatures for which $\Delta G_{\text{rxn}}^\circ$ is positive, solar process heat alone will not make the reaction proceed; some amount of high-quality energy is required in the form of work. It may be

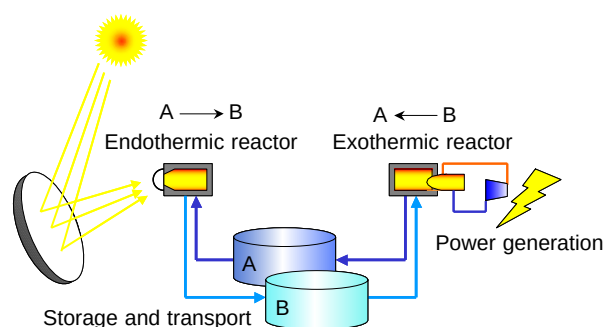


FIGURE 13 Solar chemical heat pipe for the storage and transportation of solar energy. High-temperature solar process heat is used to drive the endothermic reversible reaction $A \rightarrow B$. The product B may be stored long term and transported long range to the site where the energy is needed. At that site, the exothermic reverse reaction $B \rightarrow A$ is effected and yields high-temperature process heat in an amount equal to the stored solar energy $\Delta H_{A \rightarrow B}$. The chemical product A of the reverse reaction is returned to the solar reactor for reuse.

TABLE II

Approximate Temperatures (K) for Which $\Delta G_{\text{rxn}}^\circ$ of Reactions (26), (27), and (28) Equals Zero

Metal oxide	$\Delta G_{\text{rxn } 26}^\circ = 0$ @	$\Delta G_{\text{rxn } 27}^\circ = 0$ @	$\Delta G_{\text{rxn } 28}^\circ = 0$ @
Fe_2O_3^a	3050	920	890
Al_2O_3	>4000	2320	1770
MgO	3700	2120	1760
ZnO	2340	1220	1110
TiO_2^a	>4000	2040	1570
SiO_2^a	4500	1950	1520
CaO	>4000	2440	1970

^a Fe_2O_3 , TiO_2 , and SiO_2 decompose to lower valence oxides before complete dissociation to the metal.

supplied in the form of electrical energy in electrolytic processes or in the form of chemical energy by introducing a reducing agent in thermochemical processes. An example of a solar electrothermal reduction process that has been demonstrated experimentally is the electrolysis of ZnO using an electrolytic cell housed in a solar cavity-receiver. At 1000 K, up to 30% of the total amount of energy required to produce Zn could be supplied by solar process heat. Other interesting candidates for solar high-temperature electrolysis are MgO and Al_2O_3 .

If one wishes to decompose metal oxides thermally into their elements without the application of electrical work, a chemical reducing agent is necessary to lower the dissociation temperature. Coal (as coke) and NG (as methane) are preferred reducing agents in blast furnace processes due to their availability and relatively low price. In the presence of carbon, the uptake of oxygen by the formation of CO brings about a reduction of the oxides at much lower temperatures. Although Reactions (27) and (28) have favorable free energies above the temperatures indicated in Table II, a more detailed calculation of the chemical equilibrium composition shows that only the carbothermic reductions of Fe_2O_3 , MgO, and ZnO will result in significant free metal formation. The carbides Al_3C_4 , CaC_2 , SiC, and TiC are thermodynamically stable in an inert atmosphere; the nitrides AlN, Si_3N_4 , and TiN are stable in an N_2 atmosphere.

Using NG as a reducing agent, as indicated in Eq. (28), combines in a single process the reduction of metal oxides with the reforming of NG for the coproduction of metals and syngas. Thus, CH_4 is reformed in the absence of catalysts and, with proper optimizations, may be made to produce high-quality

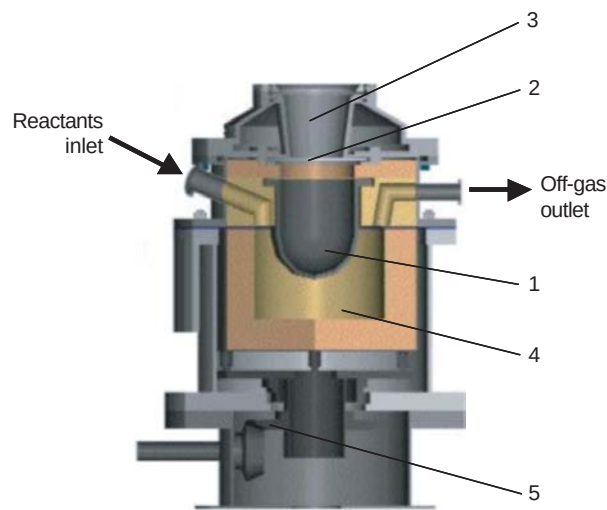


FIGURE 14 Schematic of a “two-cavity” solar chemical reactor concept for the carbothermic reduction of ZnO. It features two cavities in series, with the inner one functioning as the solar absorber and the outer one functioning as the reaction chamber. The inner cavity (1) is made of graphite and contains a windowed aperture (2) to let in concentrated solar radiation. A CPC (3) is implemented at the reactor’s aperture. The outer cavity (4) is well insulated and contains the ZnO–carbon mixture that is subjected to irradiation by the graphite absorber separating the two cavities. With this arrangement, the inner cavity protects the window against particles and condensable gases coming from the reaction chamber. Uniform distribution of continuously fed reactants is achieved by rotating the outer cavity (5). The reactor is designed specifically for beam-down incident radiation, as obtained through a Cassegrain optical configuration that makes use of a hyperbolic reflector at the top of the tower to redirect sunlight to a receiver located on the ground level. Source: Paul Scherrer Institute, Villigen, Switzerland.

syngas with an H_2/CO molar ratio of 2, which is especially suitable for synthesizing methanol—a potential substitute for petrol. Thermal reductions of Fe_3O_4 and ZnO with C(gr) and CH_4 to produce Fe, Zn, and syngas have been demonstrated in solar furnaces using fluidized bed and vortex-type reactors. These reactions are highly endothermic and proceed to completion at temperatures above approximately 1300 K. Zinc production by solar carbothermic reduction of ZnO offers a CO_2 emission reduction by a factor of 5 vis-à-vis the conventional fossil fuel-based electrolytic or imperial smelting process. Furthermore, the use of solar energy for supplying the enthalpy of the reaction upgrades the calorific value of the initial reactants by 39%.

Two examples of solar chemical reactor concepts for producing Zn by Reactions (27) and (28) are shown in Figs. 14 and 15, respectively: the “two-cavity” solar reactor based on the indirect irradiation

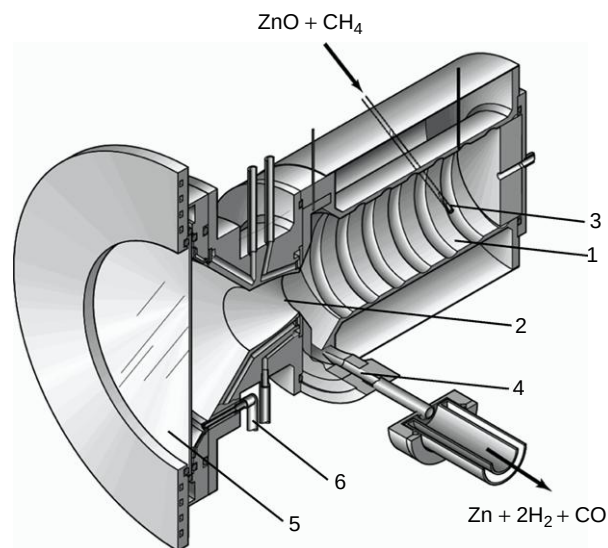


FIGURE 15 Schematic of a “vortex” solar chemical reactor concept for the combined ZnO reduction and CH₄ reforming. It consists of a cylindrical cavity (1) that contains a windowed aperture (2) to let in concentrated solar energy. Particles of ZnO, conveyed in a flow of NG, are continuously injected into the reactor’s cavity via a tangential inlet port (3). Inside the reactor’s cavity, the gas–particle stream forms a vortex flow that progresses toward the front following a helical path. The chemical products, Zn vapor and syngas, exit the cavity continuously via a tangential outlet port (4) located at the front of the cavity, behind the aperture. The window (5) is actively cooled and kept clear of particles by means of an auxiliary flow of gas (6) that is injected tangentially and radially at the window and aperture planes, respectively. Energy absorbed by the reactants is used to raise their temperature to above approximately 1300 K and to drive Reaction (28). Source: Paul Scherrer Institute, Villigen, Switzerland.

of ZnO + carbon and the “vortex” solar reactor based on the direct irradiation of ZnO + methane.

3.2 Solar Production of Material Commodities

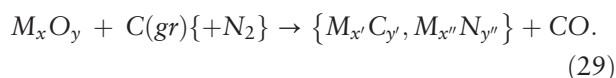
3.2.1 Solar Metals, Metallic Oxides, Carbides, and Nitrides

The extractive metallurgical industry is a major consumer of process heat and electricity and is responsible for approximately 10% of global anthropogenic greenhouse gas emissions. A solar thermal process would drastically reduce these emissions. For example, solar processing of aluminum could reduce the CO₂ emissions per ton of aluminum from 44 to 5. The temperature required to reduce aluminum ores with carbon is in the range of 2400 to 2600 K, that is, at the upper limit for practical process heat addition from a combustion source alone. Preliminary tests

performed in a concentrating solar facility suggest two feasible solar thermal routes. First, at approximately 2400 K, a mixture of Al₂O₃, SiO₂, and carbon results in an Al–Si alloy that can be purified after its production. Second, at above 2500 K, Al₂O₃ and carbon results in an Al–Al₃C₄ mixture of varying compositions.

As described in the previous subsection, Zn can be produced from ZnO by a solar thermal, electrothermal, or carbothermal reduction process. Other examples of solar thermal reduction processes that have been demonstrated in solar furnaces’ exploratory runs are the production of Zn by dissociation of its natural ore willemite (Zn₂SiO₄), the combined ZnO reduction and biomass pyrolysis, and the production of Mg and Fe by carbothermal reduction of their oxides.

Metallic carbides and nitrides can be produced by the solar carbothermic reduction of metal oxides. The reaction is generally given by



Exploratory experimental studies were conducted in solar furnaces in which nitrides (AlN, Si₃N₄, TiN, and ZrN) were formed for the systems run in N₂ and carbides (Al₄C₃, CaC₂, SiC, TiC) were formed for the systems run in inert atmosphere. These ceramics are valuable materials for high-temperature applications due to their high hardness, excellent corrosion resistance, high melting points, and low coefficients of thermal expansion. CaC₂ is well known as the feedstock for the production of acetylene. The nitrides and carbides AlN, Fe₃C, and Mn₃C may also be used in cyclic processes as feedstock to produce hydrogen and hydrocarbons or may serve as intermediaries in the production of the metal. The hydrolysis of AlN yields NH₃, the hydrolysis or acidolysis of Fe₃C yields liquid hydrocarbons, and the hydrolysis of the various carbides of manganese yields H₂ and hydrocarbons in different proportions. Thus, Reaction (29) may be incorporated into cyclic processes of the type shown in Fig. 16, in which the metal oxides that result from the hydrolysis are recycled to the solar reactor, and CO from Reaction (29) may be shifted to H₂ via the water–gas shift reaction.

3.2.2 Solar Fullerenes and Carbon Nanotubes

Fullerenes and carbon nanotubes have become a major field in condensed matter physics and chemistry. However, the large-scale production of these materials with high yield and selectivity is still a

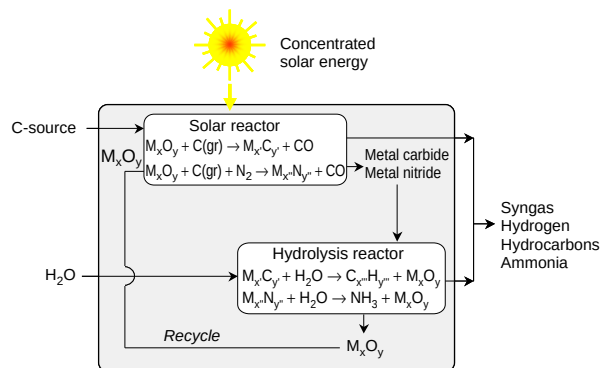


FIGURE 16 Schematic of a two-step thermochemical process for the production of syngas, hydrogen, hydrocarbons, and ammonia using metal oxides and solar energy. In the first (endothermic) step, the metal oxide is carbothermally reduced to a metal carbide or nitride using solar process heat. Subsequently, the carbide or nitride is reacted with water to produce hydrocarbons or ammonia, and CO is water-gas shifted to syngas, whereas the metal oxide is recycled to the solar step. M, metal; M_xO_y, metal oxide; M_xC_y, metal carbide; M_xN_y, metal nitride.

crucial problem. Conventional methods for the synthesis of fullerenes, such as electric arc discharge and laser ablation, fail when the process is being scaled up to higher power levels. The solar vaporization process seems to be more promising using the solar reactor shown in Fig. 17, in which a graphite rod contained under vacuum pressure behind a hemispherical quartz window is directly irradiated with peak solar concentration ratios exceeding 7000 suns and the vaporized carbon (at temperatures above 3300 K) is swept out by Ar, quenched, and collected in a filter bag. The key parameters characterizing this process are the carbon soot mass flow rate and the desired product yield. The former is a function of the target temperature, whereas the latter is a function of specific reactor variables such as fluid flow patterns, residence times, concentration of the carbon vapor in the carrier gas, target temperature, and temperature distribution in the cooling zone.

Catalytic filamentous carbon (CFC) can be produced by the solar thermal decomposition of hydrocarbons in the presence of small metal catalyst particles. Solar furnace experiments confirmed that nanotubes can be obtained using Co/MgO catalyst for CO and CH₄ + H₂ and that nanofibers can be obtained on Ni/Al₂O₃ catalyst for CO, CH₄, CH₄ + H₂, and C₄H₁₀. Figure 18 shows a solar chemical reactor used for such experimental runs. It consists of a quartz tube containing a fluidized bed of catalyst and Al₂O₃ grains. A secondary reflector, composed of a two-dimensional CPC

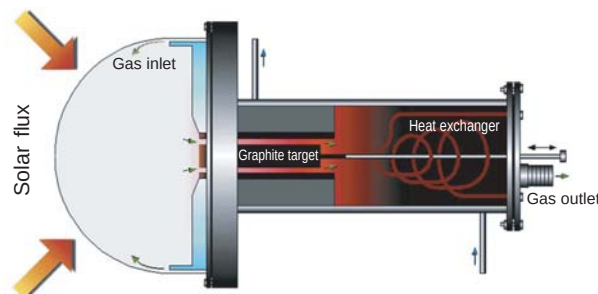


FIGURE 17 Scheme of a solar reactor for fullerene synthesis. Source: Institut de Science et de Génie des Matériaux et Prodédés, Centre National de la Recherche Scientifique, Font-Romeu, France.

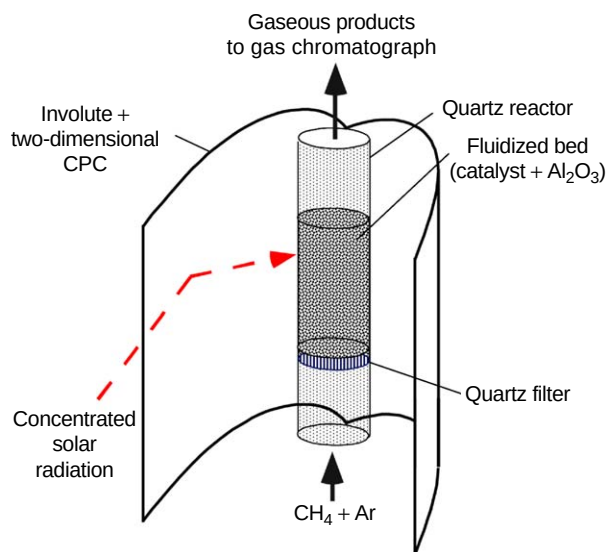


FIGURE 18 Scheme of a solar reactor for CFC production. Source: Paul Scherrer Institute, Villigen, Switzerland.

coupled to an involute, provides uniform irradiation on the tubular reactor. CFC formed typically has the following properties: surface area 100 to 170 m²/g, pore volume 0.4 to 0.8 cm³/g, micropore volume 0.004 to 0.008 cm³/g, and average pore diameter 10 to 40 nm.

3.2.3 Solar Lime and Cement

The thermal decomposition of calcium carbonate (limestone) to calcium oxide (lime),



is the main endothermic step in the production of lime and cement at 1300 K. Substituting concentrated solar energy in place of carbonaceous fuels, as the source of high-temperature process heat, is a means to reduce the dependence on conventional

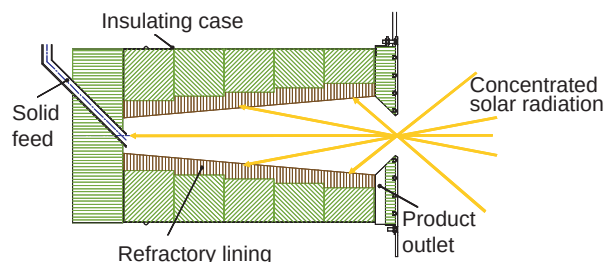


FIGURE 19 Schematic of a direct-irradiated rotary reactor for the solar production of lime. Source: Paul Scherrer Institute, Villigen, Switzerland.

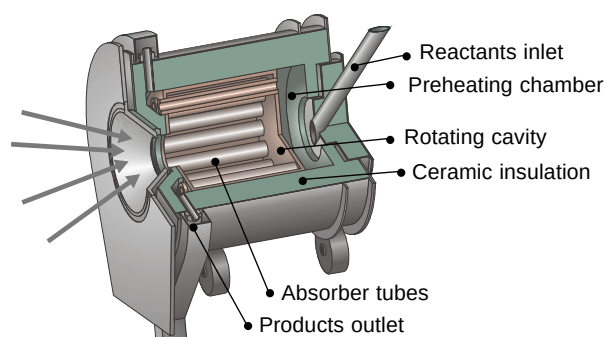


FIGURE 20 Schematic of an indirect-irradiated rotary reactor for the solar production of lime. Source: Paul Scherrer Institute, Villigen, Switzerland.

energy resources and to reduce emissions of CO_2 and other pollutants. Figures 19 and 20 show two solar reactor concepts for the solar calcination process based on direct irradiation and indirect irradiation, respectively. The direct-irradiated solar reactor (Fig. 19) consists of a refractory-lined conical rotary kiln operated in a horizontal position. Because of the rotational movement of the kiln, the reactants are transported within the conical reaction chamber from the preheating zone in the back (feeding side) to the high-temperature zone in the front (discharging side). The indirect-irradiated reactor (Fig. 20) consists of a multiple-tube absorber and a preheating chamber, both made from SiC. The well-insulated rotary reactor is tilted and works in continuous mode of operation.

3.2.4 Solar Thermal Recycling of Hazardous Waste Materials

Solid waste materials from a wide variety of sources (e.g., municipal waste incineration residuals, discharged batteries, dirty scraps, automobile shredded residue, contaminated soil, dusts, and sludge, other by-products from the metallurgical industry) contain hazardous compounds that are usually vitrified in a

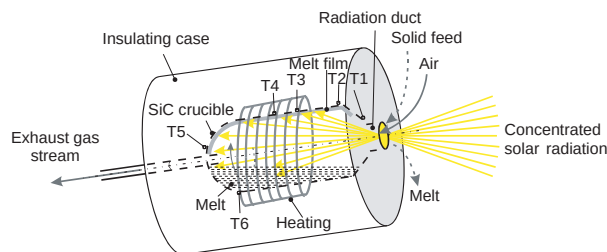


FIGURE 21 Scheme of a rotary kiln for aluminum melting. Source: Deutsches Zentrum für Luft- und Raumfahrt, Germany.

nonleaching slag and are finally disposed of at hazardous waste storage sites. However, limited storage space, increasing storage costs, and environmental regulations have led the need for developing technologies that recycle these toxic materials into useful commodities rather than deposit them in dump sites for an undetermined period of time. Thermal processes are well suited for the treatment of these complex solid waste materials. Waste materials containing carbonaceous compounds can be converted by thermal pyrolysis and gasification into syngas and hydrocarbons that can be further processed into other valuable synthetic chemicals. Those containing metal oxides may be converted by carbothermal reduction into metals, nitrides, carbides, and other metallic compounds. The commercial recycling techniques by blast, induction, arc, and plasma furnaces are major consumers of fossil fuel-based electricity and heat. Concentrated solar energy offers the possibility of converting hazardous solid waste material into valuable commodities for processes in closed and sustainable material cycles. Examples of recycling processes demonstrated in solar furnaces are the processing of electric arc furnace dust using the “two-cavity” reactor depicted in Fig. 14 and the melting of Al scrap using a SiC-lined rotary kiln shown in Fig. 21, which can be tilted for discharging molten metal and slag into molds.

Other recent applications in the area of solar thermal processing of materials that are being explored include hardening of carbon steel, production of expanded natural graphite (ENG) by exfoliation of graphite intercalated compounds, synthesis of advanced ceramics, and purification of metallurgical-grade silicon for photovoltaic applications.

3.2.5 Economics

Economic assessments indicate that solar thermochemical processes for the production of hydrogen and other solar fuels have the potential of becoming economically competitive vis-à-vis alternative paths

for producing solar fuels such as via electrolysis using solar-generated electricity. These studies further indicate that the heliostat field is responsible for approximately half of the total investment costs for the entire chemical plant, whereas the cost of the solar reactor represents 10 to 15%. However, for a fixed product throughput, the solar reactor's efficiency dictates the size of the heliostat field. Thus, reaching high solar reactor efficiencies and reducing the cost of the heliostats per unit area will have a significant impact on reducing the unit cost of the solar fuel. Application of credit for CO₂ mitigation and pollution avoidance will further enable the solar thermochemical technologies to compete favorably with fossil fuel-based processes. The weaknesses of these economic evaluations are related primarily to the uncertainties in the viable efficiencies and investment costs of the various components due to their early stage of development and their economies of scale.

4. OUTLOOK

Solar thermochemical processes have favorable long-term prospects because they avoid or reduce costs for CO₂ mitigation and pollution abatement. The

products are renewable fuels and material commodities for delivering clean and sustainable energy services. Further development and large-scale demonstration are warranted.

SEE ALSO THE FOLLOWING ARTICLES

Alternative Transportation Fuels: Contemporary Case Studies • Hydrogen Production • Materials for Solar Energy • Photosynthesis, Artificial • Solar Cells • Solar Cooling, Dehumidification, and Air-Conditioning • Solar Detoxification and Disinfection • Solar Distillation and Drying • Solar Heat Pumps • Solar Ponds • Solar Thermal Energy, Industrial Heat Applications • Solar Thermal Power Generation • Solar Water Desalination

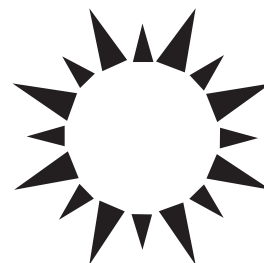
Further Reading

- Fletcher, E. A. (2001). Solarthermal processing: A review. *J. Solar Energy Eng.* 123, 63–74.
- Steinfeld, A., and Palumbo, R. (2001). Solar thermochemical process technology. In "Encyclopedia of Physical Science and Technology." (R. A. Meyers, Ed.), vol. 15, pp. 237–256. Academic Press, San Diego.

Solar Heat Pumps

KAMIL KAYGUSUZ

Karadeniz Technical University
Trabzon, Turkey



-
1. Developments
 2. Types and Applications
 3. Performance and Economics
 4. Environmental Benefits
 5. Conclusion

Glossary

air conditioning The treatment of air temperature, humidity, cleanliness, and circulation so as to achieve a controlled desired result.

coefficient of performance (COP) Ratio of heating capacity to electrical input.

compressor In a vapor compression cycle, the device that increases the pressure and temperature of refrigerant vapor; it continuously draws low-pressure refrigerant vapor from the cooler, adds energy to increase the refrigerant pressure and temperature, and discharges the high-pressure vapor to the condenser.

condenser A heat exchange coil within a mechanical refrigeration system used to reject heat from the system; it is the coil where condensation takes place.

evaporator A heat exchange coil within a mechanical refrigeration system used to absorb heat into the system; it is the coil where evaporation takes place.

heat A form of energy; it is the least organized form of energy and the form of energy into which other forms deteriorate.

heat pump A comfort system in which the refrigeration cycle is reversed by using a four-way valve to supply both heating and cooling.

thermostat A device that connects or disconnects a circuit in response to a change in the ambient temperature.

vapor compression cycle A refrigeration cycle consisting of a cooler, a compressor, a condenser, and an expansion device.

The general term heat pumping technologies is used for processes in which the natural heat flow from a higher temperature level to a lower one is reversed by adding high-value energy, that is, exergy. The term

heat pump is used in different ways in various regions of the world. In Japan and the United States, reversible units are called heat pumps; in addition to cooling and dehumidification, they offer the capability of providing useful heat. Most commonly, air-to-air units are used; however, ground-coupled systems have an increasing share. Chillers are more or less always called chillers, even if they are also used for heating purposes, as are heat pump chillers. In Europe, the term heat pump is used for heating-only units, with the heat sources outside air, ground, and groundwater, combined with heat distribution systems.

1. DEVELOPMENTS

Solar energy systems and heat pumps are two promising means of reducing the consumption of fossil energy sources and, potentially, the cost of delivered energy for domestic space heating and cooling with water heating. A logical extension of each is to try to combine the two to further reduce the cost of delivered energy. It is widely believed that solar heat pump combined systems will save energy, but what is not often known is the magnitude of the possible energy savings and the value of such savings relative to the additional expense. So, renewed interest is being directed at solar heat pumps by scientists, architects, engineers, manufacturers, and building owners as well as in the marketplace in general. The principal advantage of employing solar radiation as a heat pump heat source is that, when available, it provides heat at a higher temperature level than do other sources, resulting in an increase in the coefficient of performance (COP). Compared with a solar heating system without a heat pump, the collector efficiency and capacity are materially increased due to the lower collector temperature required.

The combination of a heat pump and solar energy system would appear to alleviate many of the disadvantages that each has when operating separately.

During winter, the energy that could be collected by the solar system, but that would be too low in temperature to be useful for direct heating, may be used as a source for the heat pump. Because the solar collector storage system can supply energy at temperatures higher than the ambient outdoor air, the capacity and COP of the heat pump would increase over those for the heat pump alone, the peak auxiliary load requirement would be reduced, and the combined heating system would seem to operate more economically. The operation of the solar system at temperatures near or below room temperature would decrease the collector losses and allow more energy to be collected. The lower collection temperature might allow the use of collectors with one or no covers, and this would reduce the first cost from a conventional two-cover solar system. Finally, for those areas where warm temperatures occur during cloudy periods, the combined system might compensate for the reduced performance of the conventional solar system under cloudy conditions and the low capacity of the heat pump in cold weather.

2. TYPES AND APPLICATIONS

Research and development in the solar-assisted heat pump (SAHP) field has been concerned with two basic types of systems: direct and indirect. In direct systems, refrigerant evaporator tubes are embodied in a solar collector, usually of the flat plate type. Research has shown that when the collector has no glass cover plates, the same collector surface can also function to extract heat from the outdoor air. The same surface may then be employed as a condenser, using outdoor air as a heat sink for cooling.

An indirect system employs another fluid, either water or air, which is circulated through the solar collector. When air is used, an air-to-air heat pump may be employed, with the collector being added in such a way that (1) the collector can serve as an outdoor air preheater, (2) the outdoor air loop can be closed so that all source heat is derived from the sun, or (3) the collector may be disconnected and the outdoor air used as the source or sink. When water is circulated through the collector, the heat pump circuit may be of either the water-to-air or water-to-water type. Heat pumps complement solar collectors and heat storage in SAHP systems in several ways:

- Heat pumps are more efficient and can provide more heat for a given heat pump size if their evaporators can operate from a warm source.

Thermal energy storage, heated by solar collectors, can provide that warming.

- Solar collectors operate more efficiently if they collect heat at lower temperatures. If the collected heat can be stored at a lower temperature because it is used to warm the evaporator of a heat pump, the collector is more efficient and, therefore, fewer collectors are needed to collect a required amount of heat.

- Heat pumps are the most efficient way in which to use electricity for backup heat for a solar heating system, even if there is no direct thermal connection of the heat pump with the solar collector and storage system.

- Heat pumps can allow sensible and latent heat storage units (e.g., water, phase change materials [PCMs]) to operate over wide temperature ranges because stored heat down to 5°C can be used in conjunction with heat pumps to heat a building.

Figure 1 shows the SAHP system linked to an energy storage tank for residential heating. In this system, solar collectors were constructed by modifying flat plate water-cooled collectors. Each solar collector had eight 1.25 cm outside diameter copper tubes spaced 10 cm apart and an effective absorber area of 1.66 m². The aluminum sheet was 0.55 mm thick. The 18 collectors were connected in parallel series form. The heat pump compressor was a hermetic type. It was driven by a 1490-W electric motor. The heat pump can use air and water as a heat source. The storage tank contained 1500 kg calcium chloride hexahydrate as a PCM for energy storage.

In general, SAHPs can work in three forms: parallel, series, and dual-source systems. A parallel system would use solar energy in the heat exchanger first and then use the air-to-air heat pump when the solar coil is unable to provide the necessary heat. A series system would use the solar energy in the heat exchanger first and then use the water-to-air heat pump when the solar coil is unable to provide the necessary heat. Dual-source heat pumps are currently in the development stage. A dual-source system would have the choice of either an air-heated coil when warm air is available or a water-heated coil when warm water is available. This type of capability would be advantageous in the SAHP, and in applications where well water is available as a source during the heating season. It would also alleviate the need for a cooling tower when the heat pump operates in the cooling mode.

It is not unusual for the SAHP to save up to 20% in operating costs during the heating season compared with an air-to-air heat pump. This represents a

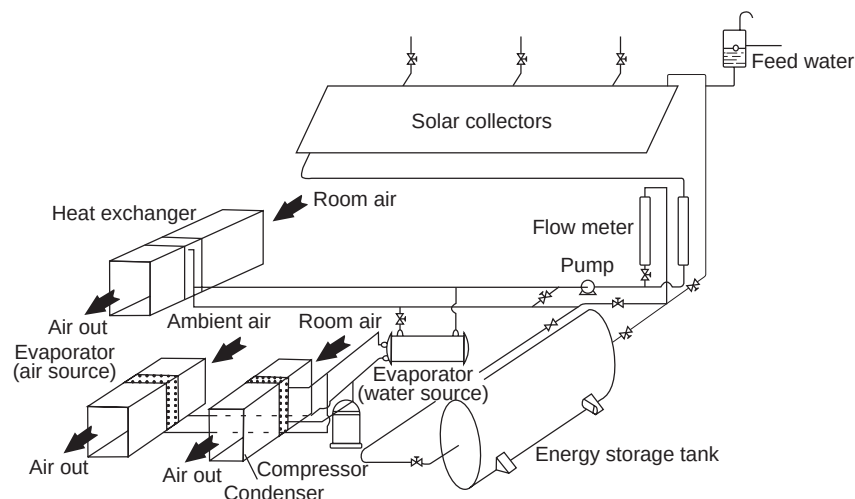


FIGURE 1 Schematic overview of experimental solar-assisted heat pump system. Reprinted from Kaygusuz *et al.* (1993).

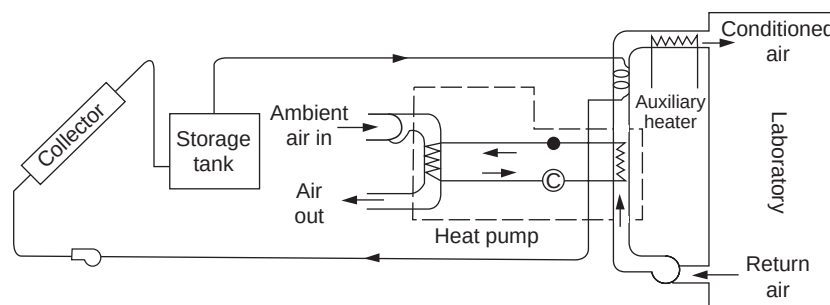


FIGURE 2 Parallel heat pump system. Reprinted from Kaygusuz *et al.* (1993).

heating COP of nearly 3.0. It has been reported that parallel systems can save up to 50% of the energy required when compared with electrical resistance heating, series systems can save up to 60% of the energy normally required, and dual-source systems can save up to 70% of the energy normally required.

A wide variety of options exist to provide air conditioning along with solar heating of buildings. The efficacy of coupling these options with a solar heating system depends strongly on which system is to be used and whether one is considering systems that can already be built reliably or ones that promise optimum performance in the future following extensive development effort. Several SAHP systems are described to illustrate their relationship with air conditioning: the simple parallel system, series system, dual-source system, and cascade system.

2.1 Parallel System

As shown in Fig. 2, the parallel system consists of a solar collector, a water-to-air heat exchanger, an air-

sourced heat pump, a water-circulating pump, a storage tank, and other equipment. The parallel SAHP system is combining two main components: the solar system and the parallel heat pump system. In this system, the heat pump uses ambient air as an energy source while the water-to-air heat exchanger uses solar energy as a heat source, and they give their energies to the load one by one. Solar energy is used to meet as much of the heating requirement as possible. Thus, the total available energy of the system is the sum of the extracted energies from two different systems (the solar system and the heat pump).

2.2 Series System

Figure 3 shows a series SAHP system. The solar collector heats water, which is then stored in a tank. The tank provides heat directly to the house if the tank temperature is above 40°C, and the heat pump draws heat from the tank when the tank temperature is between 5 and 40°C. This system has the advantage that the tank can be operated at lower

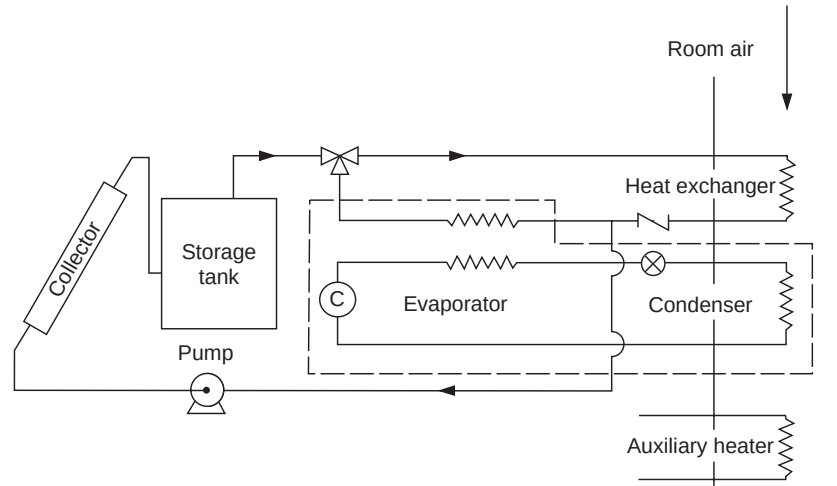


FIGURE 3 Series heat pump system. Reprinted from Kaygusuz *et al.* (1993).

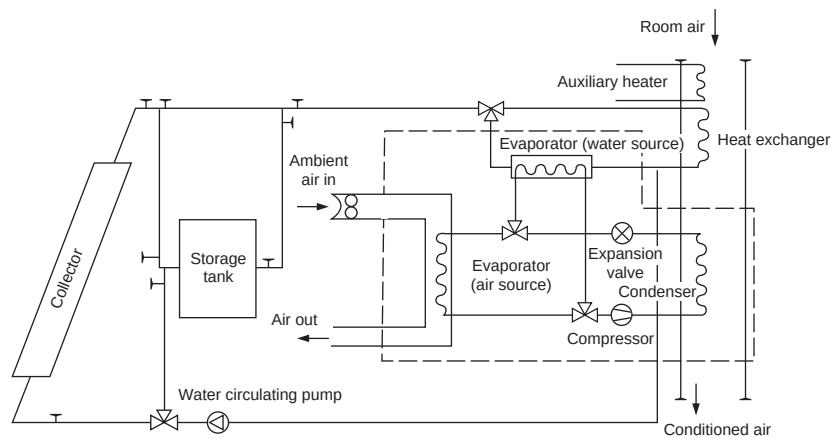


FIGURE 4 Dual-source system. Reprinted from Kaygusuz *et al.* (1993).

temperatures when required, allowing the collector to operate with high efficiency, and the tank heat storage capacity can be increased by the amount of sensible heat stored between 5 and 40°C. The disadvantage of the system is that when the tank temperature finally drops to 5°C, the heat pump cannot be used further without danger of freeze-up in the water tank. The system shown will not provide air conditioning because there is no way in which to exhaust the waste heat to outside air. It would be possible to provide air conditioning if there were another heat exchanger loop between the storage tank and outside air, but this would add to the cost of the heat exchanger and its associated circulating pump. A development effort is required on the heat pump in this system because residential heat pumps are not currently designed to operate efficiently with evaporator temperatures above 20°C, and this one would have to operate at up to 40°C.

2.3 Dual-Source System

The dual-source SAHP system, illustrated in Fig. 4, combines the advantages of the parallel and series systems and overcomes their disadvantages. The solar collector, heat storage, and building heat exchanger all continue to work as a simple solar system so long as the tank temperature remains above 40°C. At temperatures below that, the heat pump system is called on by a microprocessor controller, and a decision is made as to whether it is better to draw the heat pump's heat from the tank or from the outside air. The control strategy options are numerous for this system; it can be operated to optimize savings of electricity or to reduce peak loads. It would be the most efficient of all the systems described here. It has the disadvantage that the required heat pump is complex and will be costly to manufacture. Considerable development work will

be required to ensure that the unit can be operated in all sequences of its modes with full reliability. With this system, air conditioning is accomplished by operating the unit as a simple air-to-air heat pump in the cooling mode, using only the outside heat exchanger as a condenser. So, it is obvious that the dual-source heat pump system takes advantage of the best features of the series and parallel heat pump systems. The system is not capable of using the storage tank to reduce air-conditioning peak loads during the summer because the heat pump cannot be used to cool the tank to the outside air.

2.4 Cascade System

As Fig. 5 shows, the cascade SAHP system allows complete access to the storage tank in both the heating and air-conditioning modes. Two heat pumps are used: one between the outside air and the tank and the other between the tank and the house. The inside heat pump allows the system to operate in the series heat pump mode, providing heat to the house when the tank temperature drops below 40°C. When supplemental heat is required, the outside heat pump charges the tank with heat removed from the outside air. In air conditioning, the inside heat pump cools the house on demand, depositing the heat in storage; the outside heat pump cools the house on demand,

depositing the heat in storage; and the outside heat pump then removes heat from the tank to outside air during off-peak periods. Both of the heat pumps used in the cascade system require some development effort. The outside heat pump must be optimized for operation between outside air temperatures and the cool water temperatures of the storage tank. The inside heat pump is the same as the series unit described previously except that it must include the ability to air condition as well as heat. The required development effort would probably be less than that for a three-coil unit. This system is probably the most expensive of those described here due to the cost of heat pump equipment. When operating with the outside heat pump as a supplementary heat source, it will be less efficient than the three-coil or parallel system because the heat must go through a second heat pump in heating the house. On the other hand, it is the most flexible system, allowing full heating and air conditioning as well as the use of the storage tank to store heat for off-peak use (solar or electrical utility) during the heating season and to reduce air-conditioning peak loads during the cooling season. Today, a number of solar-assisted heat pumps are commercially available either as complete systems or as component packages, as shown in Figs 6, 7, and 8.

It is possible to view much of the work of the past few years on SAHP systems as an attempt to address a single fundamental question: what does one do when the sun is not shining and storage is depleted? Studies based on the assumption that electrical resistance is used as the backup heat source have generally concluded that the SAHP is not a good idea. However, it is now possible to identify two general paths that have been taken to avoid this difficulty: direct expansion SAHP systems and ground-coupled SAHP systems. Each type of system has its own special characteristics. These are described in what follows.

2.4.1 Direct Expansion SAHP

Refrigerant-filled, two-phase solar collectors have been used during recent years as alternatives to conventional air or water collectors. The working fluid in these collectors undergoes a liquid-to-vapor transformation, and the absorbed solar energy is manifested as the latent heat of vaporization. Two-phase collectors are usually employed in direct expansion solar-assisted heat pump (DX-SAHP) cycles. The primary objective in these cycles is to raise the evaporator temperature through solar energy input, which in turn translates into improvement in the thermal performance of the heat pump.

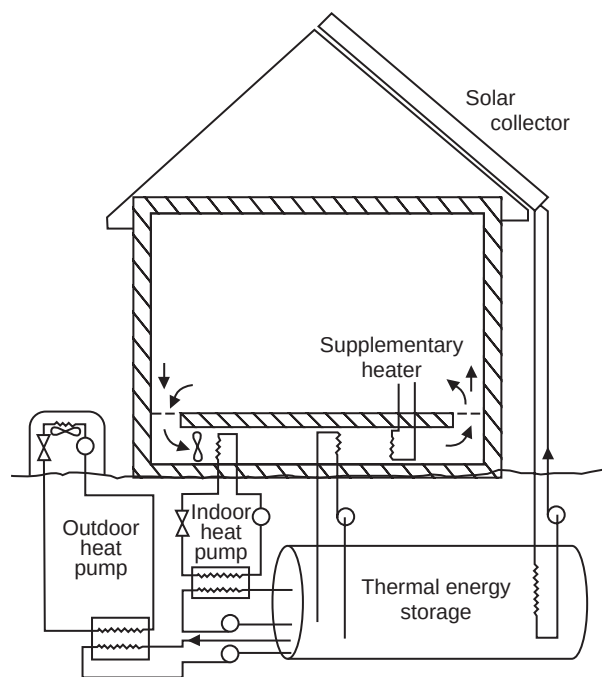


FIGURE 5 Cascade system. Courtesy of General Electric Company (reprinted from Sauer and Howell, 1991).

which uses a buried copper piping network through which refrigerant is circulated. To distinguish them from DX-GCHPs, systems using water-to-air and water-to-water heat pumps are often referred to as GCHPs with secondary solution loops.

GCHPs are further subdivided according to ground heat exchanger design: vertical and horizontal. Vertical GCHPs generally consist of two small-diameter, high-density polyethylene (PE) tubes in a vertical borehole. The tubes are thermally fused at the bottom of the bore to a close return U-bend. The nominal diameter of vertical tubes ranges from 20 to 40 mm. Bore depths range from 15 to 185 m depending on local drilling conditions and available equipment.

The advantages of vertical GCHPs are that they (1) require relatively small plots of ground, (2) are in contact with soil that varies very little in temperature and thermal properties, (3) require the smallest amount of pipe and pumping energy, and (4) can yield the most efficient GCHP system performance. The disadvantage is that they are typically higher in cost due to the limited availability of appropriate equipment and installation personnel.

Horizontal GCHPs can be divided into three subgroups: single pipe, multiple pipe, and spiral. Single-pipe horizontal GCHPs were initially placed in narrow trenches at least four deep. These designs require the greatest amount of ground area. Multiple pipes (usually two or four) placed in a single trench can reduce the required amount of ground area. With multiple-pipe GCHPs, if trench length is reduced, total pipe length must be increased to

overcome thermal interference with adjacent pipes. The spiral coil is reported to further reduce the required ground area. These horizontal ground heat exchangers are made by stretching small-diameter PE tubing from the tight coil in which it is shipped into an extended coil that can be placed vertically in a narrow trench or laid flat at the bottom of a wide trench. Required trench lengths are only 20 to 30% those for single-pipe horizontal GCHPs, but pipe lengths may be double for equivalent thermal performance.

The advantages of horizontal GCHPs are that (1) they are typically less expensive than vertical GCHPs because appropriate installation equipment is widely available, (2) many residential applications have adequate ground area, and (3) trained equipment operators are more widely available. Disadvantages include (1) the larger ground area requirement; (2) greater adverse variations in performance because ground temperatures and thermal properties fluctuate with season, rainfall, and burial depth; (3) slightly higher pumping energy requirements; and (4) lower system efficiencies.

3. PERFORMANCE AND ECONOMICS

The important quantities for the evaluation of the systems are the seasonal energy flows. For space and service hot water heating, these are the sum of the house space and water heating loads (q_{load}), the heat collected by solar energy system (q_{solar}), the heat

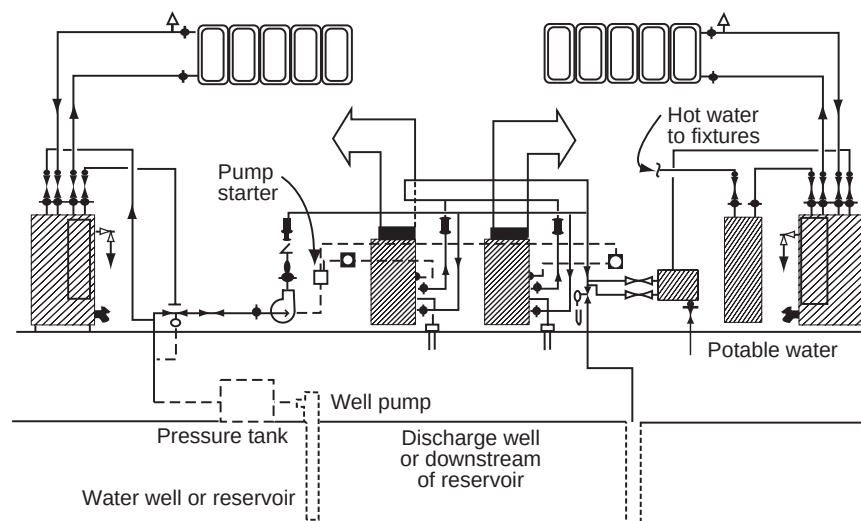


FIGURE 8 Northrup solar-assisted heat pump system and components. Courtesy of Northrup Inc. (reprinted from Sauer and Howell, 1991).

extracted by the heat pump from the ambient air (q_{air}), the electricity required by the heat pump compressor and fans (W_{hp}), and the auxiliary energy required to meet the hot water and space heating loads (E_{aux}). Here, E_{aux} is the delivered energy and is less than the purchased fuel energy due to furnace and water heater inefficiencies. Over the course of a year, there is a negligible change in the amount of energy in the storage tank, and the system energy balance is

$$q_{\text{solar}} + q_{\text{air}} + W_{\text{hp}} + E_{\text{aux}} = q_{\text{load}}.$$

The first two terms are free energy, whereas the second two terms represent purchased energy. A measure of the thermal performance of combined solar heat pump systems is the fraction (F) of the total load that is met by nonpurchased energy, defined as

$$F = (q_{\text{air}} + q_{\text{solar}})/q_{\text{load}} = 1 - (W_{\text{hp}} + E_{\text{aux}})/q_{\text{load}}.$$

The auxiliary energy may be supplied by gas, oil, electricity, or a combination, whereas the heat pump work is nearly always electricity. The thermal performance fraction does not distinguish among the different kinds and prices of fuels; rather, it reflects only energy requirements.

The fraction, F , is shown in Fig. 9 for Madison as a function of collector area for the conventional furnace, conventional heat pump, parallel, series, dual-source, and conventional single-cover solar system. For a house with neither a solar energy system nor a heat pump, F equals zero. For a house with only a heat pump, the fraction of the heating

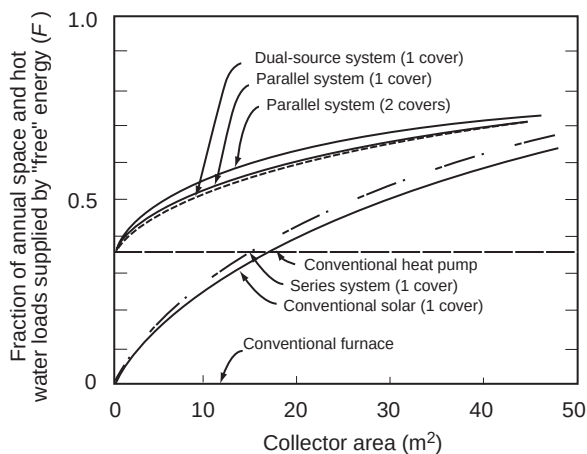


FIGURE 9 The fraction of energy from nonpurchased source (F) for series, parallel, dual-source, heat pump-only, and standard solar energy systems as a function of collector area, based on simulations of the systems on a residential building in the Madison climate. Reprinted from Freeman *et al.* (1979).

requirement supplied by nonpurchased energy is q_{air} divided by the total heating requirement and equals 36%. Because the heat pump does not contribute to the domestic hot water system, the value of F depends on the COP and the relative sizes of the space and water heating loads and is usually between 0.2 and 0.4. For a house with a conventional solar system, F depends on collector area. As collector area increases, F increases from zero asymptotically toward 1.0.

The results of the simulations done using the program TRNSYS show that with the same collector area, the parallel system is substantially better than the series system and slightly better than the dual-source system in all collector sizes in that it delivers a greater fraction of the loads from nonpurchased sources. This arises because the heat pumps in the series and dual-source systems must operate to deliver all solar energy stored below 20°C. The extra electrical energy required to deliver this energy more than compensates for the combined advantages of higher collector efficiency and higher heat pump COP. The temperature of storage in the series system (and in the dual-source system) in these simulations is rarely high enough to permit direct solar heating.

Improved collector performance and higher q_{solar} not only are the direct consequences of adding the solar source capability but also are the only benefits to system performance during the heating season. However, the improved collector performance is achieved at the expense of heat pump electrical power input that usually displaces a cheaper fuel. The annual energy balance requires that the sum of all energy supplied equals the load given as

$$q_{\text{load}} = q_{\text{solar}} + q_{\text{air}} + W_{\text{hp}} + E_{\text{aux}}.$$

The relative contributions from each of these four heat sources are shown in the bar graphs in Fig. 10. The combined height of the q_{solar} and q_{air} bars in the figure represents the percentage of the total heating requirement supplied by free energy and, therefore, is equivalent to F . The systems are ranked from left to right in order of decreasing savings of purchase energy that is decreasing F .

On the other hand, the experimental heat pump seasonal COP (i.e., ratio of rejected heat to work input) varies among the systems. The use of a solar source for the heat pump raises the seasonal COP over that of the conventional heat pump and parallel systems. As shown in Fig. 10 for Trabzon, Turkey, the annual heat pump heating COPs for the parallel, series, and dual-source systems are 3.0, 4.0, and 3.5, respectively. The relative COP values for the

conventional heat pump and the parallel system heat pump result from the fact that the heat pump is forced to operate less often but under generally more unfavorable conditions when used as an auxiliary to a solar system. As expected, the COP for the series and dual-source heat pumps are substantially higher because they use the solar-heated storage tank as a source. Table I shows the experimental performance of the SAHP system for residential heating in Trabzon.

Two configurations of series SAHPs were evaluated experimentally in TECH House I on the University of Tennessee campus in Knoxville. During the 1979–1980 heating season, a 4.2-m³ insulated above-ground storage tank was used. During the 1980–1981 heating season, the storage system consisted of an uninsulated 7.6-m³ steel tank buried

2.1 m in the ground. For 1979–1980, the heating season system performance factor was 2.42 and the system's maximum peak electric demand was 11.20 kW. For 1980–1981, the heating season performance factor increased to 3.16, whereas the maximum peak electric demand decreased to 4.33 kW. On the other hand, a parallel SAHP system was evaluated experimentally in TECH House IV for three heating seasons between the fall of 1979 and the spring of 1982. The seasonal performance factors (SPFs) for the 1980–1981 and 1981–1982 seasons were measured to be 2.56 and 2.11, respectively.

In the competition for commercial viability, all solar heating systems must compete for some version of life cycle economic advantage with alternative nonsolar heating systems. In going to solar heating, an economic trade-off is made between the high first cost of solar heating or cooling equipment and the cost savings realized by reduced use of fuel or electricity. One way in which to compare nonsolar and solar heating systems is to determine the cost of electricity at which they have the same overall life cycle cost. A rough comparison of this break-even cost of electricity shows that many solar-assisted systems will be more cost effective than the conventional air-to-air heat pump available in the United States today only when electricity costs become greater than 10 to 12 cents/kWh. The current national average cost of residential electricity is approximately 4 cents/kWh, with a range of up to approximately 10 cents/kWh in some major urban utility systems.

The feasibility of SAHP systems for space heating and domestic hot water preheating in Canada is

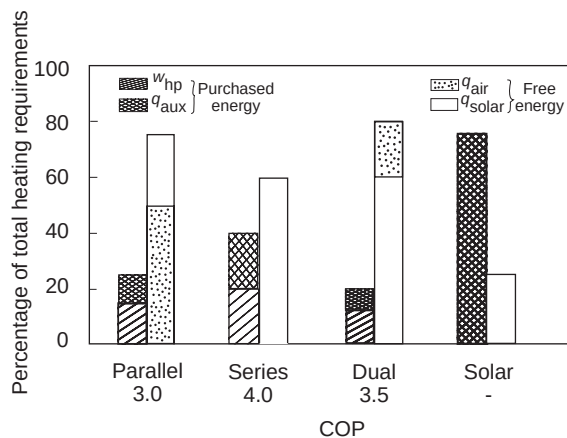


FIGURE 10 Heating contributions from all possible sources for Trabzon, Turkey. Reprinted from Kaygusuz (1995).

TABLE I

The Experimental Performance of the Solar-Assisted Heat Pump System for Trabzon, Turkey

Heating month	N	Parallel system				Series system			
		COP	D_{col} (percentage)	D_{st} (percentage)	F (percentage)	COP	D_{col} (percentage)	D_{st} (percentage)	F (percentage)
November	30	2.90	50	56	80	21	4.40	60	60
December	31	2.92	52	58	76	13	4.30	60	45
January	26	2.80	53	62	48	15	4.27	61	48
February	23	2.70	48	60	38	20	4.20	62	50
March	30	2.85	52	63	75	22	4.32	64	64
April	30	3.00	55	64	85	23	4.40	68	70
May	10	3.10	60	66	90	7	4.50	66	80

Source. Kaygusuz (1999a, 1999b).

Note. N, number of working days of the heat pump system per month; F , fraction of the load met by free energy; COP, coefficient of performance; n_{col} , collector efficiency; n_{st} , storage efficiency.

examined by simulating the performance of these systems on a computer using the program WATSUN. Simulations are carried out using meteorological data for seven representative Canadian cities, two different building types, and six types of system configurations. For the SAHP system, 20-year life cycle cost comparisons are made with two reference systems, a conventional resistance heating system and an air-to-air heat pump system, based on current economic parameters and projected escalation scenarios for electricity rates.

Results of the study show that the SAHP systems conserve significant amounts of energy over resistance heating and heat pump systems. On the life cycle unit cost basis, SAHP system costs are relatively insensitive to location, but the dependence on building types is substantial, with multiplex dwellings showing the least cost. Liquid-based, dual-source SAHP systems are found to be cost-effective over resistance heating at some of the locations for multiplex units. They are not cost-effective for single-family dwellings at the current time.

4. ENVIRONMENTAL BENEFITS

Energy efficiency and environmental benefit are the key drivers for heat pump application and further research and development. A simple projection of world energy consumption and dwindling energy sources during the coming decades shows that society faces a major challenge. There is a general awareness of this rapidly approaching problem, but society as a whole does not seem to be responding accordingly, assuming that there is sufficient time to find solutions. However, it is encouraging to see that there is a growing interest in applying energy-efficient technology in developing countries and economies in transition. Although costs are an obstacle for applying modern technology, the industrialized world seems increasingly prepared to look for new ways of information and technology transfer to these developing parts of the world.

How important are heat pumps in relation to the environment? Any thermodynamic specialist or heating/cooling engineer will have no problem in quickly answering this question. The numbers on COPs and SPFs, as well as on specific carbon dioxide (CO_2) emission levels, are available and can lead to only one conclusion: an increasing application of heat pumps without delay is very important. On the other hand, it is frequently argued that the refrigerants used in heat pumps today are safe and that their impact on the

environment during the equipment life cycle is negligible. That may be true and significant as well, but the fact remains that a much better job can be done. There is significant scope for improved performance and initial cost reduction. Both will lead to a better environment and decreased energy consumption.

In Fig. 11, various systems for the supply with electricity and heat are compared with regard to their CO_2 emissions. The various lines in the figure represent standard systems that are already installed or at least are obtainable in Turkey. The figure shows the CO_2 emissions in kilograms for the supply with 1 kWh power plus heat (in kilowatt-hours). On the other hand, in the case of a conventional coal-fired power plant with an electrical efficiency of 0.40 and a distribution efficiency of 0.95, 0.9 kg CO_2 is emitted for each kilowatt-hour of power delivered to the user. All systems in the figure use this power plant, so all curves intersect at the same point on the ordinate. The steepest line represents a system combining this coal power plant with an oil furnace for heating. For example, for the supply with 1 kWh power from the power plant and 2 kWh heat from the furnace, we get CO_2 emissions of 1.7 kg. When we change the fuel for the furnace from oil to natural gas, we get a 12% reduction in CO_2 , from 1.7 to 1.5 kg, for this example. If we further improve the boiler efficiency up to 1.0, which is the value for modern condensing furnaces, we obtain a reduction of the CO_2 emissions of 24%.

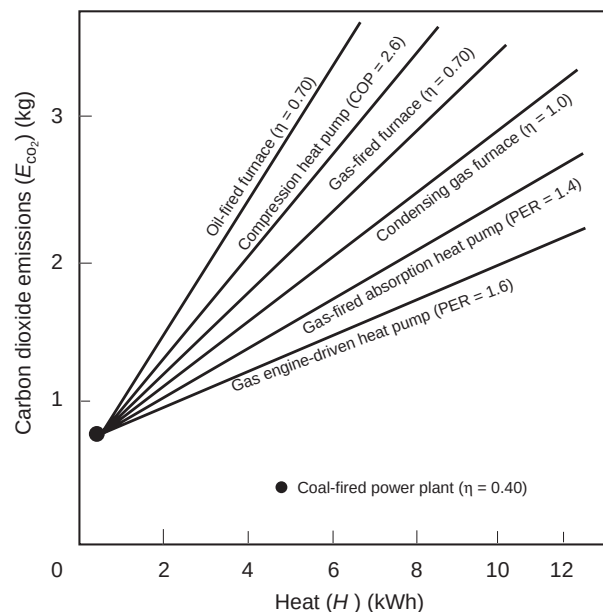


FIGURE 11 Carbon dioxide (CO_2) emissions for systems with a coal-fired power plant. Reprinted from Kaygusuz (2003).

Compression heat pumps with a COP of 2.5, operated with electric power from coal power stations, show only slightly less emissions than do oil furnaces and are even worse than standard gas furnaces and especially condensing gas furnaces. On the other hand, for a larger heating demand, only gas engine-driven compression heat pumps, with a primary energy rate (PER) of 1.5, or direct gas-fired absorption heat pumps (PER = 1.3) show a similarly high reduction in CO₂ emissions. The efficiencies assumed in Fig. 11 are not the best values for available compression heat pumps. Modern systems reach a COP of 3.0 and sometimes 4.0. Modern power stations with a combined gas-and-steam turbine cycle (CC) reach electrical efficiencies of up to 0.53.

In Fig. 12, the CO₂ emissions are plotted for these systems. The striped areas indicate the range of COPs of the advanced heat pumps. As can be clearly seen, heat pumps driven by electricity from conventional coal power stations cannot reduce the CO₂ emissions in comparison with condensing gas furnaces in spite of the COP improvement from 2.5 to 4.0. The situation changes appreciably with modern CC power plants with an efficiency of 0.53. In such a system, even heat pumps with a rather poor COP of 2.5 have lower emission ratios than do the best furnaces. Of course, much better results are achieved with COPs of up to 4.0, which are also shown in the figure.

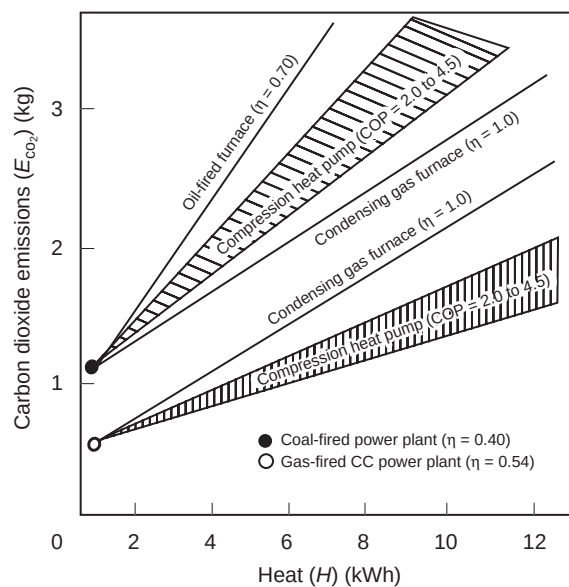


FIGURE 12 Carbon dioxide (CO₂) emissions for systems with advanced compression heat pumps. Reprinted from Kaygusuz (2003).

Worldwide, more than 100 million heat pumps are now installed. Their total annual thermal output is approximately 1300 TWh. This represents an annual equivalent CO₂ emissions reduction of 0.13 billion tons (0.6%), against the backdrop of annual global emissions of CO₂ of more than 22 billion tons. However, the current reduction potential of heat pumps is 6% of global CO₂ emissions, and improved technology could increase this to 16% in the future.

5. CONCLUSION

Heat pumps use mechanical energy to transfer thermal energy from a source at a lower temperature to a sink at a higher temperature. Electrically driven heat pump heating systems have attracted wide interest as an alternative to electric resistance heating or expensive fuels. They have two advantages: a COP greater than unity for heating (thereby saving on purchase of energy) and usefulness for air conditioning during the summer. Heat pumps may use air or water sources for energy, and dual-source machines that can use either are under development.

SAHPs for residential applications have received a great deal of attention since 1975. They combine a solar heating system with a heat pump to provide the necessary heating and cooling requirements. Both air-to-air and water-to-air heat pumps can be used in this system. The solar system basically collects the energy for the heating requirements, and the heat pump supplies this energy at the proper time and temperature and also helps to improve the efficiency of the solar collector in the water-to-air operating mode.

On the other hand, heat pump systems have other advantages for commercial buildings besides their low operating costs. They offer improved comfort due to on-demand, year-round heating or cooling. Greater flexibility in the interior layout of the building is also possible, particularly where a distributed water loop configuration is adopted. In addition, more exterior architectural freedom is possible with the adoption of GSHP systems due to the absence of any need for external equipment such as cooling towers. Because of their use of electric power, heat pumps can also provide a convenient way in which to heat a building in a remote location to which fossil fuel is not economically transportable.

In the past, heat pumps have often been a tool of government policy. For example, during the oil crises of the mid-1970s and early 1980s, governments in some cold climate countries encouraged the use of heat pumps through financial incentives to reduce

their dependence on oil and as part of their energy conservation programs. This may happen again as countries sign on to the Kyoto Protocol and as governments implement measures to reduce greenhouse gas emissions.

As a result, heat pumps are capital intensive, as are solar energy systems. Consideration of the economics of SAHP systems indicates that it may be difficult to justify the investment in a heat pump to improve solar energy system performance unless the cost of the heat pump can be justified on the basis of its use for air conditioning or unless the use of the heat pump would reduce collector costs to a small fraction of their current levels. Today, the most popular commercial heat pump configuration in the United States, Japan, and other countries such as Turkey is the single-packaged air source unit. This is followed by the packaged terminal heat pump and the water source heat pump.

SEE ALSO THE FOLLOWING ARTICLES

Ground-Source Heat Pumps • Materials for Solar Energy • Solar Cells • Solar Cookers • Solar Cooling, Dehumidification, and Air-Conditioning • Solar Detoxification and Disinfection • Solar Distillation and Drying • Solar Ponds • Solar Thermal Energy, Industrial Heat Applications • Solar Thermal Power Generation • Solar Water Desalination

Further Reading

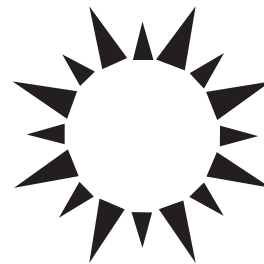
American Society of Heating, Refrigerating and Air-Conditioning Engineers. (1995). Geothermal energy. In "ASHRAE

- Handbook: HVAC Applications," chap. 29. ASHRAE, Atlanta, GA.
- Badescu, V. (2002). Model of a solar-assisted heat-pump system for space heating integrating a thermal energy storage unit. *Energy Buildings* 34, 715–726.
- Bouma, J. (2002). Heat pumps: Better by nature—Report on the Seventh IEA Heat Pump Conference. *IEA Heat Pump Centre Newsletter* 20(2), 10–27.
- Chandrashekar, M., Le, N. T., Sullivan, H. F., and Hollands, K. G. T. (1982). A comparative study of solar assisted heat pump systems for Canadian locations. *Solar Energy* 28, 217–226.
- Chaturvedi, S. K., Chen, D. T., and Kheireddine, A. (1997). Thermal performance of a variable capacity direct expansion solar-assisted heat pump. *Energy Convers. Mgmt.* 39, 181–191.
- Duffie, J. A., and Beckman, W. A. (1991). "Solar Engineering of Thermal Processes." John Wiley, New York.
- Freeman, T. L., Mitchell, J. W., and Audit, T. E. (1979). Performance of combined solar-heat pump systems. *Solar Energy* 22, 125–135.
- Kavanaugh, S. (1998). Ground source heat pumps. *ASHRAE J.* 40(12), 31–36.
- Kaygusuz, K. (1995). Performance of solar-assisted heat-pump systems. *Applied Energy* 51, 93–109.
- Kaygusuz, K. (1999a). Investigation of a combined solar-heat pump system for residential heating: I Experimental results. *Intl. J. Energy Res.* 23, 1213–1223.
- Kaygusuz, K. (1999b). Investigation of a combined solar-heat pump system for residential heating: II. Simulation results. *Intl. J. Energy Res.* 23, 1225–1237.
- Kaygusuz, K. (2003). "Environmental aspects of fossil fuel combustion and reducing greenhouse gas emissions with heat pump systems in Turkey." Unpublished manuscript, Karadeniz Technical University, Trabzon, Turkey.
- Kaygusuz, K., Gültekin, N., and Ayhan, T. (1993). Solar-assisted heat pump and energy storage for domestic heating in Turkey. *Energy Convers. Mgmt.* 34, 335–346.
- O'Dell, M. P., Mitchell, J. W., and Beckman, W. A. (1984). Design method and performance of heat pumps with refrigerant-filled solar collectors. *ASME J. Solar Energy Eng.* 106, 159–164.
- Reid, R. L., McGraw, B. A., and Bedinger, A. F. G. (1985). Performance of series solar-assisted heat pump systems in TECH House I. *ASME J. Solar Energy Eng.* 107, 293–296.
- Sauer, H. J., and Howell, R. H. (1991). "Heat Pump Systems." Krieger Publishing, Malabar, FL.

Solar Ponds

N. D. KAUSHIKA

Indian Institute of Technology
New Delhi, India



-
1. Introduction
 2. Solar Pond Taxonomy
 3. Salt Gradient Solar Pond
 4. Saltless Nonconvective Solar Ponds
 5. Applications
 6. Economics
 7. Environmental Impact

Glossary

aquifer A layer of soil or rock that can hold water.
double-diffusive fluid A fluid embodying the density gradients of two constituents.
nonconvective layer A fluid layer in which natural convection is suppressed.
permeable Allows liquid or gas to pass through.
solar pond Body of still water that collects solar energy and stores it as thermal energy.
thermohaline system Solar thermal system.
transparent insulation layer A layer that is transparent for solar (optical) radiations and is opaque for infrared (heat) radiations.

The solar pond represents a low-cost option among other solar systems. It possesses thermal storage capability spanning seasons and is the only viable system with 100% solar fraction. Solar pond heat can be used for a variety of applications, including generation of electricity, industrial process heating, desalination, and greenhouse heating.

1. INTRODUCTION

Solar radiant energy is characterized by low-energy density and its practical utilization involves large-area solar collectors. A solar pond is a unique large-area solar collector that uses a body of still water

to collect solar energy and stores it as thermal energy. As a practical matter, all ponds, lakes, oceans, and other expanses of water in nature collect solar energy and convert it to thermal energy, but their heat retention efficiency is poor because the water near the surface is quickly cooled as the heat is rapidly dissipated to the environment. The warmer and more buoyant water in the lower region rises to the surface; this movement of water is called natural convection. Therefore, a water-pond system will be more effective in the collection and storage of solar energy if the thermal losses to the environment are reduced either by suppressing the natural convection of water in the pond or by covering the pond surface with a partially transparent insulation. The term solar pond is therefore used for a water pool with reduced heat losses. It can collect solar energy, concentrate and store it as heat up to approximately 100°C (higher temperature restricted by the boiling point of water). The solar pond involves simple technology and uses water as working material for three functions: collection of solar radiant energy and its conversion into heat, storage of heat, and transport of thermal energy out of the system.

2. SOLAR POND TAXONOMY

The first recorded use of solar energy for heating a water pool was made approximately 2500 years ago when great Roman baths were heated by the sun. Von Kálcinsky, who reported that the bottom of Madve lagoon in Transylvania in summer had a temperature of approximately 70°C, discovered the phenomenon of the solar pond in 1902. In this lake, salt concentration was also observed to increase with depth. The downward increase in salt concentration prevents the convection and renders the upper region of the lake a partially transparent insulator trapping solar heat in the bottom region. The

conceptual design of the first artificial “sun pool” was given by Tabor and Bloch in 1958. The proposed system was an artificial pond (approximately 1 m deep) with a black bottom and salt gradient with a maximum concentration at the bottom obtained by filling the pond with progressively less dense brine layers. The name salt gradient solar pond or solar pond was coined for such a system. The term solar pond has since been applied to a variety of water pool solar collector/storage systems. These may be broadly classified as shown in Fig. 1.

In nonconvective solar ponds, the basic principle is that heat loss to the environment is reduced by the suppression of natural convection in the upper layer (transparent insulation layer) of water or air medium. The most documented type of nonconvective solar pond is the salt gradient solar pond. It uses such salts as NaCl, MgCl_2 , and bitterns. It is referred to as an unsaturated salt-stabilized pond to distinguish it from the saturated salt-stabilized pond based on the Soret effect. The latter uses salt such as KNO_3 , whose solubility increases dramatically with temperature; in the pond, saturation of the salt concentration is maintained at all depths. The pond is hotter in the bottom region than at the top. Thus, a progressively larger amount of salt is dissolved toward the bottom. Owing to saturation at each level, the vertical diffusion of salt is checked and the density gradient is stable. This provides for the

possibility of a maintenance-free solar pond. Alternative methods of suppressing convection have also been investigated and several solar pond systems have been proposed and tested, including the gel- or viscosity-stabilized pond, the membrane pond, and the honeycomb-stabilized solar pond. These are referred to as saltless nonconvective solar ponds.

In convective solar ponds, heat loss to the environment is reduced by covering the pond surface. The cover is transparent and the pond is often of shallow depth; it is thus referred to as a shallow solar pond. In practice, it is a water-filled plastic bag glazed at the top with a blackened bottom insulated by foam. These ponds are used for producing hot water at medium temperature ($40\text{--}60^\circ\text{C}$) for industrial applications as well as for boiler preheating. Figure 2 provides a schematic of solar pond concepts.

Earliest theoretical as well as experimental studies of solar ponds predicted solar heat collection efficiencies of approximately 20% and thermal storage capability spanning seasons. Several countries have since built and operated salt gradient solar ponds. The technology of salt gradient ponds is inherently simple. However, for a solar pond energy system of good thermodynamic efficiency, a variety of factors must be considered and quantified. Therefore, the physics and engineering aspects of salt gradient solar ponds and the conceptual outline of other ponds are presented in the following sections.

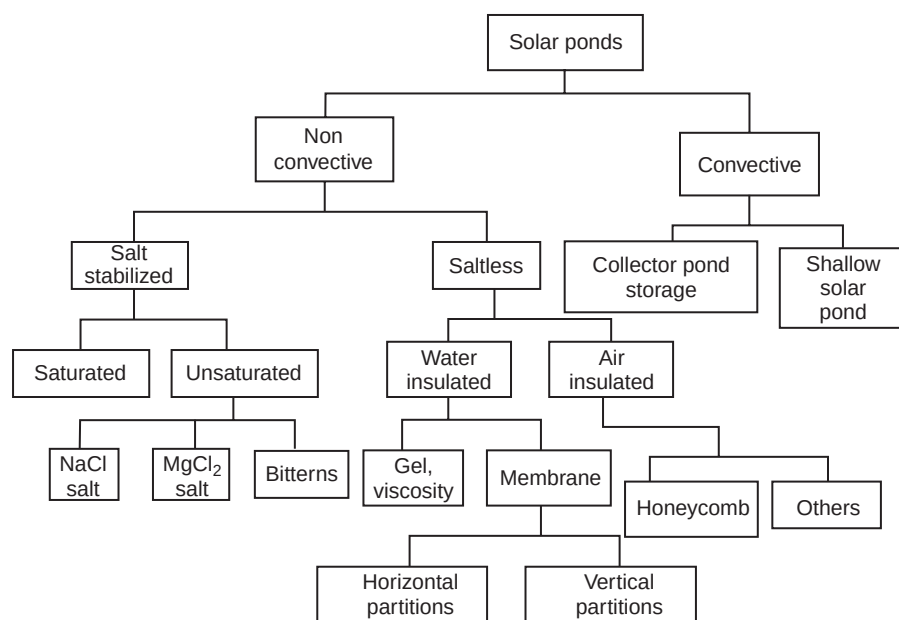


FIGURE 1 Classification of solar ponds.

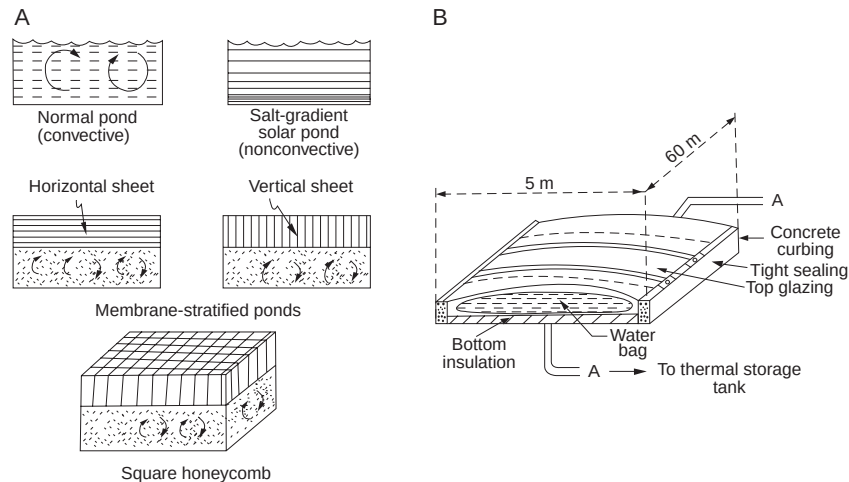


FIGURE 2 (A) Schematics of nonconvective solar pond concepts. (B) Illustration of shallow (convective) solar pond.

3. SALT GRADIENT SOLAR POND

3.1 Containment

Solar pond containment involves trade-offs between several strategies related to the location of the pond, earth excavation, the lining, insulation, and pond shape. Local geology is critically important in site selection; care must be taken to avoid underground aquifers. The pond is usually contained in earth excavated to 3 or 4 m depth. The excavation may be minimized by stacking earth to make a wall around the pond. Leakage effects from the pond can cause considerable loss of salt solution as well as stored heat. In case of permeable soil, these effects may be quite severe; care must be taken when lining the bottom and sides. A membrane liner of butynol, nylux solid vinyl, ethylene propylene diene monomer, or other suitable synthetic materials may be used. An alternative to liner is to treat the soil in such a way as to make it impermeable to hot brine solution. For example, an inexpensive liner may be buried under a layer of compacted earth.

Thermal insulation in the pond bottom and sides may be used to reduce heat losses; it may speed up pond heating but at excessive cost. Pond shape is also an important consideration. Circular surface shapes provide a minimum perimeter for a given pond surface area, involve minimum heat losses to ground, and hence are thermally the most efficient. Rectangular shapes do not have this advantage but are usually less expensive to construct. Slopping side-walls are very effective in reducing heat loss. Sloppy walls are also desirable from the standpoint of

structural stability of banks and edges; a gradient of 45° has generally been used.

3.2 Establishment and Control of Gradient

Three methods have been adopted for the establishment of initial salt density gradient in solar ponds: natural diffusion, stacking, and redistribution. The first method relies on the natural diffusion between a freshwater layer and the layer of saturated salt solution. The pond is half filled with each layer and the surface and bottom concentrations are maintained constant. Stacking involves the filling of a pond with layers of progressively changing concentrations. The redistribution method is considered to be most convenient and expedient for large-area solar ponds. In practice, the redistribution method involves filling the pond with brine of near-saturation concentration (S_c) to half of its total depth (L); the fresh water is then fed into the pond through the diffuser, which is initially placed at the bottom in the brine. The water will stir and uniformly dilute the pond brine above the diffuser. The pond surface level will start rising, and the diffuser is simultaneously raised upward in a continuous motion or in discrete steps of up to 5 cm. The timing of diffuser movement is adjusted so that the diffuser as well as the brine surface reach the final level at the same line. Consequently, in the pond region progressively diluted brine layers are generated and a constant (S/L) density gradient, positive downward, is established.

The mathematical expression for the rate of diffuser movement was derived by Zangrando as follows: If $x_s(t)$ and $x_d(t)$ represent the height of the pond surface and the diffuser with reference to the pond bottom at any time t , then in accordance with the proportionality criteria, we have

$$x_d(t) = b_1 x_s(t) + b_2, \quad (1)$$

where b_1 and b_2 are constants:

$$\text{At } t = 0, x_d(t) = 0 \text{ and } x_s(t) = 1/2 \quad (2)$$

$$\text{At } t = t_f, x_d(t) \text{ and } x_s(t) = 1. \quad (3)$$

Equation (1) in combination with Eqs. (2) and (3) yields

$$x_d(t) = 2x_s(t) - 1$$

or

$$\dot{x}_d(t) = 2\dot{x}_s(t). \quad (4)$$

Equation (4) implies that the rate of diffuser motion should be twice that of the surface motion.

The solar pond may be considered as a horizontal layer of incompressible fluid of depth l and infinite horizontal dimension (y, z plane); the x -axis is taken along the vertical direction and positive downward. The layer embodies the gradients of temperature and salinity corresponding to cold water at the top and hot brine at the bottom. Owing to gradients, diffusion of heat and salt occurs continuously. The solar pond may therefore be considered a thermohaline double-diffusive fluid system, and its physical behavior may be described by the following equations of fluid mechanics and salt diffusion:

1. Conservation of mass (continuity equation):

$$\frac{\partial \rho}{\partial t} + \nabla(\rho \vec{V}) = 0$$

or

$$\frac{D\rho}{Dt} + \rho \nabla \vec{V} = 0$$

2. Conservation of momentum (Navier–Stokes equation):

$$\rho \frac{D\vec{V}}{Dt} = -\nabla p + F + \mu \nabla^2 \vec{V} + \left(\xi + \frac{\mu}{3} \right) (\nabla \vec{V})$$

3. Conservation of energy (heat conduction equation):

$$\frac{DT}{Dt} = \alpha_T \nabla^2 T + Q_T$$

4. Conservation of salt:

$$\frac{DS}{Dt} = \alpha_S \nabla^2 S + Q_S,$$

where Q_T is the heat source term resulting from the absorption of solar radiation in the pond, Q_S is the salt source term resulting from the process of salt replenishment in the pond,

$$\vec{V} = \begin{bmatrix} u \\ v \\ w \end{bmatrix}$$

is the convective velocity vector, $\frac{D}{Dt}$ is the convective derivative given by

$$\frac{D}{Dt} = \frac{\partial}{\partial t} + \vec{v} \nabla,$$

μ is the coefficient of viscosity, F is the external force, and ξ is the viscosity coefficient.

Due to thermal expansion, the fluid of the bottom region tends to be lighter than the fluid at the top. Thus, the system has a top-heavy configuration, which is potentially unstable. The stability criteria in such a situation have been obtained as follows:

$$\left(\frac{\partial S}{\partial X} \right)_{\min} = \frac{v + \alpha_T}{v + \alpha_S} \times \frac{\alpha_c}{\beta_c} \left(\frac{\partial T}{\partial X} \right)$$

or

$$\left(\frac{\partial S}{\partial X} \right)_{\min} = \frac{p_r + 1}{p_r + \tau_s} \times \frac{\alpha_c}{\beta_c} \left(\frac{\partial T}{\partial X} \right), \quad (5)$$

where $p_r = v/\alpha_T$ is the Prandtl number, and $\tau_s = \alpha_S/\alpha_T$ is the ratio of diffusivities and referred to as the Schmidt number. Equation (5) gives the minimum salt gradient requirement corresponding to a given temperature gradient in the double-diffusive solar pond. This equation is often referred to as the hydrodynamic stability criterion in the context of static stability criteria based on a simple buoyancy model given as

$$\left(\frac{\partial S}{\partial X} \right)_{\min} = \frac{\alpha_c}{\beta_c} \left(\frac{\partial T}{\partial X} \right), \quad (6)$$

where α_c is thermal expansion coefficient, and β_c is the salt expansion coefficient.

For dilute solutions at ambient air temperature, $v = 7\alpha_T \approx 10^3 \alpha_S$. Therefore,

$$\frac{p_r + 1}{p_r + \tau_s} \approx 1.14.$$

This implies that a concentration gradient approximately 14% greater than the static stability is required by the hydrodynamic stability criterion.

The vertical gradient of density causes an upward diffusion of salt in the pond. For a typical sodium chloride solar pond, the amount of upward diffusion of salt is 60–80 g/m² day. The tendency of salt diffusion is to destroy the gradient. One gradient

control approach is to remove salt from the surface and inject it back into the bottom. For this approach, an evaporation pond is required. However, a more practical approach, which does not require the continual addition of salt, is the “falling pond” concept. A schematic of the falling pond technique is shown in Fig. 3. Hot brine is withdrawn from the bottom layer without disturbing the layers above. This is possible since in a fluid system stratified with a density gradient, selective flow of the bottom layer can be accomplished in the same manner as the withdrawal of a single card from the deck. The hot brine withdrawn from the solar pond is passed through the flash evaporator to remove some of its water. The solution, which now has a higher concentration and a smaller volume, is reinserted into the pond bottom. Consequently, the concentration gradient is maintained. The decrease in surface level is restored by the addition of fresh water to keep both the pond depth and the surface concentration constant. In the preceding process, the gradient tends to be displaced downward, hence the name falling pond technique. Following Tabor, if v_0 is the vertical velocity (measured positively downward) of the mass of the brine in the pond, V_1 and V_2 are values of volume flow rates at input and output, and A is the horizontal area of the pond (assumed to be uniform vertically), then

$$v_0 A = V_2 - V_1.$$

The resultant downward flux of salt is given by

$$v_0 S = (V_2 - V_1)S/A.$$

The condition for the maintenance of the gradient may be written as

$$(V_2 - V_1)S/A = \alpha_s(\partial S/\partial x),$$

where α_s is the coefficient of salt diffusion.

3.3 Wind and Precipitation Effects

Wind interaction with a solar pond generates waves and currents in the surface layer. Waves cascade energy through a wave spectrum that results in turbulence, which in turn depletes the salt concentration gradient and can promote complete convection in the surface region. The surface currents interact with banks and edges and create circulations in the vertical plane and cause layer mixing. The raindrops also considerably influence the dynamics of gradient erosion in the pond surface region. The wind flow over the pond surface increases the convective as well as evaporative heat losses. Suppression of wind effects is therefore very important for increasing the operating efficiency of solar ponds. Various methods of limiting the wind mixing have been tested or proposed, including floating plastic pipes, floating plastic nets, floating

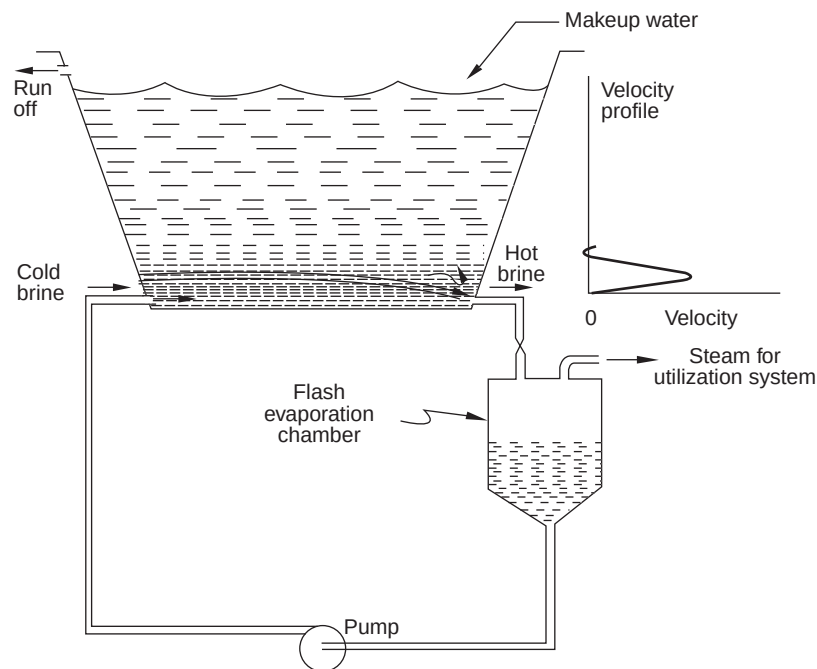


FIGURE 3 Schematic of falling pond technique of gradient control and heat removal.

plastic rings, wind baffles, and an immiscible liquid (amyl acetate and silicone oil) surface layer.

3.4 Fouling and Corrosion

Fouling and corrosion are inherent problems in solar ponds. Fouling may be caused by the falling of wind-blown material such as dirt and leaves as well as by the growth of algae or bacteria in the body of the pond. Heavy wind-blown debris normally settles to the pond bottom, where its effect on pond performance is not significant. The material of low-specific gravity is usually confined to the convective layer at the pond surface and can be removed by filtering and skimming the layer. Biological growth can be checked by adding a biocide such as copper sulfate with salt at the time of filling the pond. A hot and saline environment is usually the cause of corrosion, which is often encountered in feed water pumps and heat exchangers. Careful material selection can help. Bronze, ceramic, stainless steel, and plastics (polypropylene) are well suited for saline and hot water.

3.5 Thermal Performance

In practice, a salt gradient solar pond is a three-zone configuration. The temperature gradients are relatively large in the surface region where most of the infrared portion of the solar spectrum is absorbed; the gradient induces instabilities and results in the formation of the surface convective zone. Various processes, including wind mixing, penetrative convection, and diffusive action of salt and heat, also contribute to the formation of this zone. In normal conditions, its depth ranges between 0.2 and 0.5 m. This layer is a liability on solar pond heat collection and storage since in a pond of fixed depth the growth of the mixed surface layer takes place at the expense of the thickness of the insulating gradient layer. The mixed surface layer does not provide thermal insulation to the storage layer and absorbs a considerable portion of incident solar energy. Similarly, the excessive absorption (up to 30%) of radiation by the pond floor induces thermohydrodynamic instabilities in the bottom region and makes it convective and isothermal. The bottom convective zone provides thermal storage. A convective storage layer is therefore placed in the bottom of all modern solar ponds. The minimum depth of this zone should be 0.5 m, which is sufficient to level the diurnal fluctuations in the storage zone temperature. Depths of more than 1.25 m are very effective in leveling the seasonal fluctuations, and at a depth of 10 m no

seasonal variation is observed. A storage layer 1.2 m deep is usually considered appropriate. The optimum depth of the nonconvective zone is a function of collection temperature, which seems to favor a depth of approximately 1.25 m for this zone.

The following are typical thermal performance efficiencies of solar ponds:

Predicted: 20–30% at a collection temperature of 70–90°C

Achieved: 15% (Israel ponds) and 9–12% (U.S. ponds)

4. SALTLESS NONCONVECTIVE SOLAR PONDS

A saltless nonconvective solar pond was conceived primarily to overcome numerous problems encountered with conventional salt gradient solar pond. It consists of a convective water reservoir insulated on the bottom and sides. The surface insulation to the water is provided by a nonconvective layer of transparent fluid such as ethanol and silicone oil. The convective stability of the insulation layer is characterized by the stability of the fluid layer with adverse temperature gradient. In hydrodynamics, this problem is referred to as the Rayleigh–Bennard problem. The maximum stable temperature gradient across the layer is given by

$$(dT/dx)_{\max} = \frac{R_c K \nu}{g \beta D^4},$$

where g is acceleration due to gravity, β is the coefficient of volumetric expansion, dT/dx is the adverse temperature gradient, D is the depth of the layer, K is the thermometric conductivity or thermal diffusivity of fluid, ν is the kinematic viscosity of the fluid, and R_c is the critical Rayleigh number.

The insulation layer is nonconvective if its characteristics (ν , D , and R_c) are matched to the temperature difference between its bottom and top surfaces. Shaffer suggested that the insulation layer could be stabilized against convection by using viscous fluids such as silicon oil and water syrups. The control of convection by an array of horizontal membranes has been investigated by Hull. It has been shown that the separation of membranes (D) must be approximately 3–5 mm for convection to be suppressed; hence, to minimize the reflection losses at the interface, it is necessary to match the refractive index of fluid to that of the membrane. Water and ethanol are recommended as liquids for the surface

insulation zone. Teflon film seems to be an appropriate membrane. Ethanol is preferable since it can float on the hot water reservoir.

Another method of convection suppression in the insulation layer (air) involves the use of honeycomb panels. It consists of a sealed air-filled honeycomb panel floating over the hot water reservoir. It is predicted to be an attractive alternative, giving solar efficiencies of 40–60% at collection temperatures of 70–90°C. This is approximately double the efficiency of conventional salt gradient ponds.

5. APPLICATIONS

5.1 Electric Power Conversion

The Solar pond is ideal for electricity generation. One remarkable advantage is that its upper surface is cool due to evaporation and can be used for cooling the condenser. Thus, a power-generation application pond can accomplish three functions: radiant energy collection, thermal energy concentration and storage, and heat rejection. The heat available from salt gradient solar ponds is at temperatures of 50–100°C. Thus, the basic element involved in electricity generation is the low-temperature turbine. Based on the nature of the turbine, two plant types are possible: flashed steam plants and binary (organic) fluid cycle plants. Commercial turbines that operate with flashed steam cycle conditions are not available.

However, binary fluid cycle turbines are commercially available. These turbines are similar to steam turbines and are based on the Rankine power cycle. Their working fluid is not steam but an organic fluid, such as freons, toluene, chlorinated hydrocarbon, or hydrocarbons with a low boiling temperature. These are high-molecular-weight fluids and so the turbine is smaller than the corresponding steam turbine. These turbines obtain 0.4–0.7 Carnot efficiency. A conceptual schematic of a binary fluid cycle-based solar pond power plant is shown in Fig. 4. Israel is the world leader in the operation of such plants. A 150-kW electric power plant was put into operation at Ein Bokek, Israel, in 1979.

5.2 Process Heating

Solar pond heat holds great promise for application for process heating. Salt production is one of the earliest applications of the solar pond. Any of the following industries and industrial processes may be supplied heat from a solar pond; salt and mineral production; drying and curing of timber; milk pasteurization; concentration and separation by evaporation; cleaning and washing in the food industry; tanning of skins; textile processing, such as wool scouring, carbonizing, and dyeing; industrial laundry; and the paper industry for preheating. Solar pond heat has been demonstrated to be applicable to the social sectors of both the developed and the developing world. Specific applications

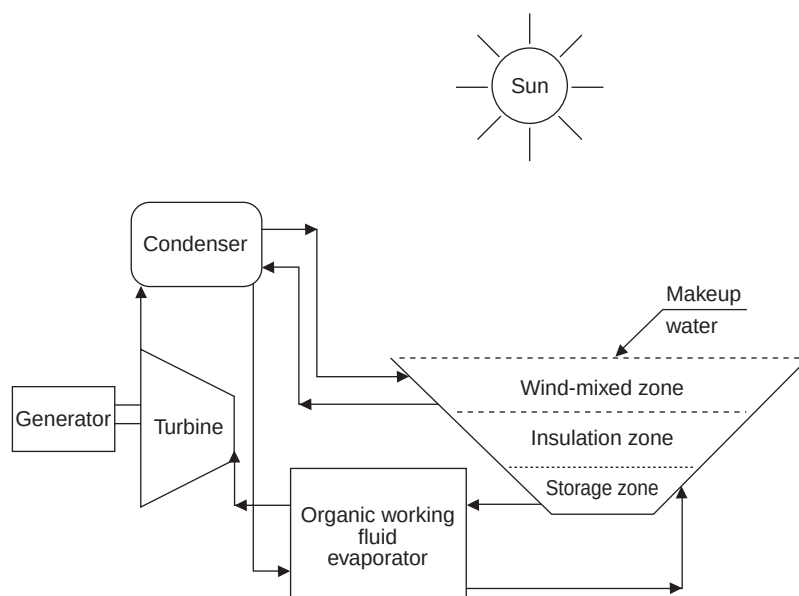


FIGURE 4 Solar pond power system.

include heating swimming pools, low-temperature cooking in southern India and western Africa, and desalinating water in outback Australia. The solar pond heat has multiple applications in the agricultural industry. At the Ohio Agricultural Research and Development Center (OARDC), a solar pond ($18.5 \times 8.5 \times 3.0$ m) has been constructed to supply the winter heat requirements of a single-family residence and partial needs of an adjacent greenhouse. In this system, hot brine is extracted from the pond and passed through shell and tube heat exchangers. If the pond temperature is higher than 40°C , the extracted heat is supplied to a water-to-air discharge heat exchanger in the green house; if the pond temperature is less than 40°C , extracted heat is first upgraded by a heat pump and then supplied for the heating of greenhouse. The method is quite suitable for this application and has good thermodynamic efficiency.

6. ECONOMICS

The major cost factors of a solar pond system are the earthwork (excavation, leveling, and compaction), the synthetic lining, the salt, the water, the land, and other miscellaneous supplies. These factors vary considerably with the location and the size of the system. Consequently, small ponds are considerably more expensive than large ponds, and there is a wide variability in the per unit area cost estimates of solar ponds operated at different locations throughout the world. Installation costs of two small-size solar ponds are provided in Table I. These estimates exclude all expenses related to the heat recovery systems.

TABLE I
Cost of Small-Size Pond Installation

Item	Australian pond (\$/m ²) ^a	American pond (\$/m ²) ^b
Earth work	9.81	5.00
Liner	2.79	11.00
Salt	1.91	9.70
Water	0.025	—
Land	—	—
Miscellaneous	0.52	5.90
Total	15.06	31.60

^a Location: Alice Springs (Sheridan, 1982).

^b Location: Miamisburg (Wittenberg and Harris, 1981).

7. ENVIRONMENTAL IMPACT

Stabilization of the greenhouse effect is the focus of global environmental concerns. The carbon dioxide emissions resulting from the use of fossil fuels must be reduced by more than 60% to stabilize the greenhouse gases at present concentrations. Stabilization cannot be achieved simply by improving fossil fuel technologies and demand management. The world, in which process heat and electricity are the major needs, has therefore pinned its hopes on the introduction of renewable energy systems. The solar pond is an inherently simple and inexpensive renewable energy means of heat and electricity. It therefore has the greatest potential for cost-effective implementation of renewable energy in the short term and long term. However, in a salt gradient pond approximately one-third of a ton of salt per unit area is used and at certain locations salt contamination may cause serious environmental hazards, such as the detrimental effect of salt in drinking water and toxicity to the vegetation surface by saline soil. The development of the honeycomb solar pond eliminates the expense of large quantities of salt and the potential problems of degrading the aquifer with salt. Fouling resulting from the falling of wind-blown debris is another inherent problem in solar ponds.

SEE ALSO THE FOLLOWING ARTICLES

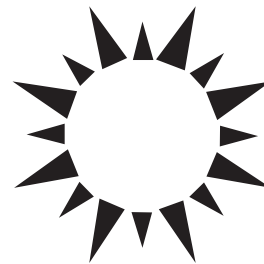
Materials for Solar Energy • Solar Cells • Solar Cookers • Solar Cooling, Dehumidification, and Air-Conditioning • Solar Detoxification and Disinfection • Solar Distillation and Drying • Solar Fuels and Materials • Solar Heat Pumps • Solar Thermal Energy, Industrial Heat Applications • Solar Thermal Power Generation • Solar Water Desalination • Thermal Energy Storage

Further Reading

- Clark, A. E., and Dickinson, W. C. (1980). Shallow solar ponds. In "Solar Energy Technology Handbook" (W. C. Dickinson and P. N. Cheremisinoff, Eds.), p. 377. Dekker, New York.
- Hull, J. R. (1980). Membrane stratified solar ponds. *Solar Energy* 25, 317.
- Hull, J. R., Nielsen, C. E., and Golding, P. (1989). "Salinity-Gradient Solar Ponds." CRC Press, Boca Raton, FL.
- Kaushika, N. D. (1986). Solar ponds. *Adv. Energy Systems Technol.* 5, 75-165.
- Nielsen, C. E. (1980). Non-convective salt-gradient solar ponds. In "Solar Energy Handbook" (W. C. Dickinson and P. N. Cheremisinoff, Eds.). Dekker, New York.

- Ortabasi, U., Dyksterhuis, F., and Kaushika, N. D. (1983). Honeycomb stabilized saltless solar pond. *Solar Energy* 31, 229.
- Roothmeyer, M. (1980). Soret effect and salt-gradient solar ponds. *Solar Energy* 25, 567.
- Shaffer, L. H. (1978). Viscosity stabilized solar pond. In "Proceedings of the ISES Congress, New Delhi, India."
- Sheridan, N. R. (1982). Solar salt pond at Alice Springs, Report 1382. Queensland University, Queensland, Australia.
- Tabor, H. (1980). Nonconvecting solar ponds. *Philos. Trans. R. Soc. London Ser. A* 295, 423.
- Tabor, H., and Weinberger, H. Z. (1980). Nonconvecting solar ponds. In "Solar Energy Handbook" (J. F. Kreider and F. Kreith, Eds.). McGraw-Hill, New York.
- Weinberger, H. (1964). The physics of solar ponds. *Solar Energy* 8, 45.
- Wittenberg, L. J., and Harris, M. J. (1981). Construction and startup performance of the Miamisburg salt-gradient solar pond. *J. Solar Energy Eng.* 103, 11.
- Zangrando, F. (1980). A simple method to establish salt-gradient solar ponds. *Solar Energy* 25, 467.

Solar Thermal Energy, Industrial Heat Applications



SANJAY VIJAYARAGHAVAN and D. Y. GOSWAMI

University of Florida

Gainesville, Florida, United States

-
1. Solar Thermal Collection Systems
 2. Solar Industrial Process Heat Overview
 3. Examples of SIPH Systems
 4. Design and Performance of SIPH Systems
 5. Economics of SIPH

Glossary

concentrator A solar collector that uses optical methods to concentrate solar radiation before absorption.

flat plate collector A solar collector that uses a flat plate geometry as the absorbing surface.

insolation A term that stands for incident solar radiation.

PTC Abbreviation for parabolic trough collector.

solar industrial process heat (SIPH) Process heat supplied using solar energy.

solar load fraction The fraction of the thermal load that is met by the solar system.

TRNSYS (Transient System Simulation) A widely used transient system simulation program developed at the University of Wisconsin.

utilizability Fraction of solar flux absorbed by the collector that is delivered to the working fluid.

Heat energy forms a large fraction of the total energy consumption in the industrial sector. For example, processes such as boiling, distillation, and polymerization that require heat input are common in chemical industries. Some other examples include cleaning and sterilization operations in the food industry, processes involving drying of materials, and the production and purification of metals and alloys. Heat is often supplied in the form of hot water and steam, or hot air, and sometimes involves the direct heating of the components in the process. All heat energy used in this fashion is referred to as industrial process heat (IPH). Most of the IPH demand today is

met by burning fuels such as natural gas, oil, and coal or by using electric heating. Solar energy can be used to supply the heat input required in IPH systems. Several such systems have been designed and constructed all over the world. With the proper choice of collection system, solar systems can be built to meet any temperature level that is required by an industrial process.

1. SOLAR THERMAL COLLECTION SYSTEMS

1.1 Solar Collectors

The simplest version of a solar collector is the flat plate collector. A typical flat plate collector consists of an absorber plate that absorbs well over the range of the solar radiation spectrum. Glazing is used to reduce heat losses by thermal convection and radiation while keeping the absorber plate transparent to solar radiation. Thermal insulation is also used to reduce losses. Flat plate collectors have a working temperature of less than 70°C. To achieve higher temperatures, evacuated tube collectors, which essentially have a vacuum between the absorber and glazing tube, can be used. Because of the very reduced conduction and convection losses across the vacuum, these collectors yield higher efficiency at lower temperatures.

Concentrating collectors use optical concentration to increase the intensity of solar radiation and, thereby, achieve higher temperatures. The simplest concentrating collector can be made by using flat wall reflectors to concentrate solar radiation on a flat plate collector. Concentrating collectors can be of tracking and nontracking types. A tracking concentrator has its orientation adjusted to account for the variation in

the position of the sun in the sky with time. For temperatures up to 350°C, parabolic trough collectors (PTCs) are used. These collectors focus solar radiation on a line focus where the absorber is located. A reflecting spherical or paraboloidal bowl is used when temperatures on the order of 250 to 500°C are needed. For extremely high temperatures, a large field of tracking mirrors (called heliostats) is used to concentrate solar radiation on a receiver that is located on top of a central tower.

1.2 Solar Ponds

A solar pond combines collector and energy storage in one unit. Solar radiation entering a pond travels some distance before it is absorbed in the water, increasing the temperature of the water at that depth. In a salt gradient solar pond, the density of water is increased at the lower section of the pond by the addition of a salt into the pond. Although the temperature of the lower section is increased by the absorption of solar radiation, the convection of warmer water to the surface of the pond is suppressed due to the density gradient imposed by the added salt. Salts such as magnesium chloride (MgCl_2) and sodium chloride (NaCl) are commonly used in solar pond applications. Bottom-layer temperatures on the order of 70 to 80°C are achievable.

1.3 Thermal Storage

Solar energy is a resource that is not available during the night and is only intermittently available on cloudy or overcast days. Many solar systems incorporate thermal storage to compensate for the intermittent nature of the solar resource. Thermal energy is normally stored as sensible heat, latent heat, or the heat of chemical reaction. In sensible heat storage, the energy is used to raise the temperature of a storage medium. Commonly used sensible heat storage materials include water, organic oils, molten salts, ceramics, and rocks. Latent heat storage takes advantage of the large amount of energy required to effect phase transformations in materials. Phase change occurs at a fixed temperature called the transition temperature or over a very small temperature range. Materials such as paraffin waxes, inorganic salts, and eutectics of organic and/or inorganic compounds are used as phase change materials. Thermal energy can also be stored as the heat of reaction in a reversible chemical reaction. The amount of energy stored depends on the heat of reaction and the extent of conversion achieved. An

advantage of thermochemical energy storage is that the products of reaction can be stored at room temperature. Examples of reactions include decomposition of metal hydrides, oxides, peroxides, and ammoniated salts.

2. SOLAR INDUSTRIAL PROCESS HEAT OVERVIEW

It is estimated that solar thermal technology could theoretically supply 51% of the U.S. industrial process heating needs. Process heat solar collectors could be used throughout the year, and each system could be designed to fit the temperature level required for its specific applications (this is particularly important in the use of process heat). Table I shows the heat energy consumption by various SIC (Standard Industrial Classification) categories for selected industries in the United States. A majority

TABLE I

Summary of U.S. Industrial Heat Use by SIC Category for 1971 and 1994

SIC group	Quantity in 10 ¹² kJ	
	1971	1994
20 Food and kindred products	779	1254
21 Tobacco products	14	W
22 Textile mills	268	327
23 Apparel	22	W
24 Lumber and wood products	188	518
25 Furniture	37	73
26 Paper and allied products	2006	2812
27 Printing and publishing	16	118
28 Chemicals	2536	5621
29 Petroleum products	2576	6688
30 Rubber	158	303
31 Leather	19	W
32 Stone, clay, and glass	1541	996
33 Primary metals	3468	2597
34 Fabricated metal products	295	387
35 Machinery	283	260
36 Electrical equipment	213	256
37 Transportation	310	383
38 Instruments	53	113
39 Miscellaneous	72	W
Total	14,854	22,854

Note. W, withheld to avoid disclosing data for individual establishments.

of the heat is used in mining, food, textiles, lumber, paper, chemicals, petroleum products, stone-clay-glass, and primary metals industries. In addition to the quantity of heat, the quality (i.e., temperature) is very important to match the proper collection system to the application.

Solar collection and concentrator technologies are capable of providing a broad range of temperatures to match the demand of various processes. Fig. 1 gives the distribution of the U.S. process heat

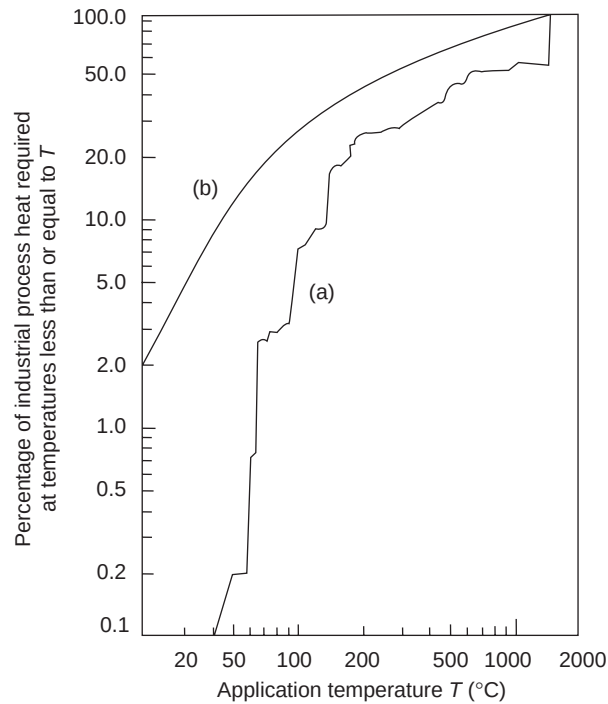


FIGURE 1 Distribution of U.S. process heat use by required temperature level: (a) heat requirements; (b) IPH requirements plus preheat from 15°C.

requirements based on the required temperature level. Note that approximately 60% of all U.S. process heat requirements fall below 370°C. This temperature is easily achievable by using PTCs. It is estimated that energy demand in the European Union for processes at a temperature below 250°C is 1080×10^{12} kJ. Table II shows the temperatures that can be achieved with low- and medium-temperature collectors and lists a few examples of potential applications.

An example of a solar industrial process heat (SIPH) system for providing process hot water or steam is shown in Fig. 2. Solar energy is collected by circulating a heating fluid (e.g., oil, water) through the solar collectors. The choice of the collector type is strongly influenced by the temperature requirements of the process and the choices made at the time of design. For instance, in the Gould solar system (described later), parabolic troughs were used to collect solar energy, with oil circulated inside the collectors to remove the heat. Because the cost per unit collector area is very dependent on the optics used in the collection process, matching the collector characteristics to temperature requirements is very important.

The hot fluid leaving the collector is stored in a storage tank. This hot fluid can then be used to supply heat to the industrial process by exchanging heat with the process fluid. Because solar energy is an intermittent resource, most applications require a backup energy source to ensure that the requirements of the industry are met even on cloudy days and days with insufficient sunshine.

Another way in which to compensate for the intermittent nature of the solar resource is to use a storage subsystem. Using storage is particularly useful if the load is of an intermittent nature as well given

TABLE II

List of Some Low- and Medium-Temperature Solar Collector Geometries and Potential Industrial Applications

Type of collector	Typical working temperature range (°C)	Example applications	
		Industry	Use, energy form
Flat plate collector	≤ 70	Food industry	Washing, hot water
		Automobile manufacturing	Paint drying, hot air
		Chemicals	Borax drying, hot air
High-efficiency flat plate collector	60–120	Lumber	Log conditioning, hot water
			Kiln drying, hot air
		Dairy	Sterilization, steam
Parabolic trough collector	150–350	Textiles	Finishing, steam
		Paper and pulp	Kraft pulping, steam

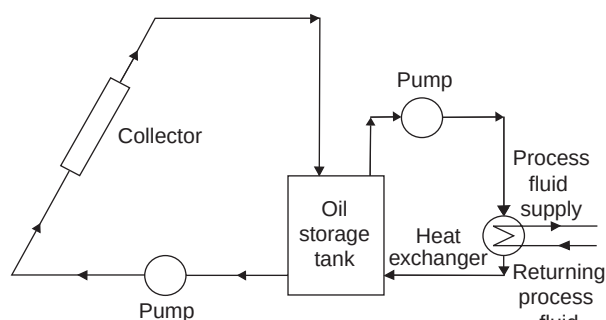


FIGURE 2 Simplified schematic of a process fluid supply system.

that the stored energy can be used to ensure that the solar system is used better. Typically, industrial process heat (IPH) systems tend not to use large storage. Instead, the solar system meets a fraction of the load, and the difference is made up using a backup source. Some SIPH systems are intentionally designed to meet only part of the load of the application. They may simply serve as the preheating arrangement so that the load on the conventional system may be reduced, thereby saving energy.

3. EXAMPLES OF SIPH SYSTEMS

Most of the IPH systems below 200°C require hot water, steam, or hot air. If the heat is needed in process air, water heating collectors along with water to air heat exchangers may be used or air heating collectors and rock storage may be used. Hot air typically is needed in agricultural drying, which may be provided by passive solar air heaters. If a large body of water is available, solar ponds may be used for low-temperature applications.

A high-temperature SIPH system may be designed to use low-, medium-, and high-temperature collectors in stages so as to minimize the cost and maximize the efficiency. A schematic of such a system designed for textile dyeing process is shown in Fig. 3.

3.1 SIPH for Textile Industries

Of all the energy used in the textile industry, 60 to 65% is used in wet processing, including dyeing, finishing, drying, and curing. The energy for wet processing is used as hot water and steam. The textile industry in the United States uses approximately 500 billion L of water per day, and approximately 25% of this water is used at an average temperature of 60°C.

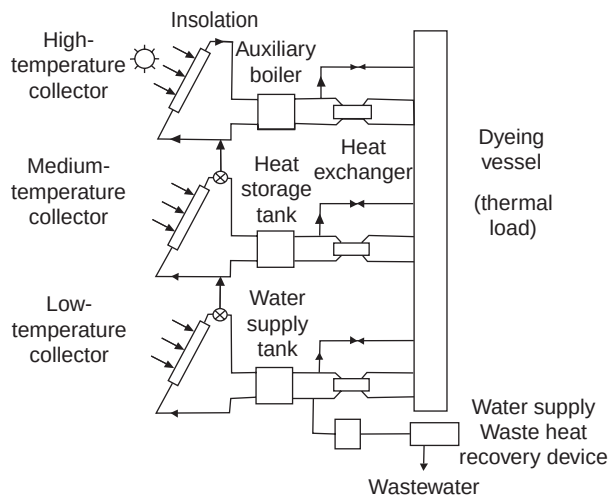


FIGURE 3 Schematic of a SIPH system for a high-temperature dyeing process.

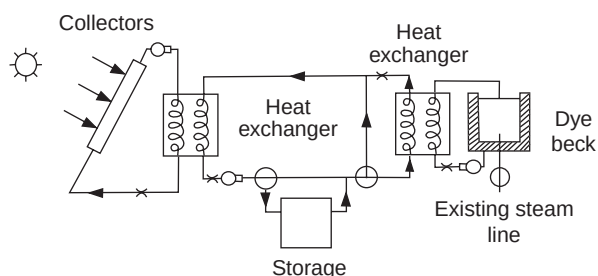


FIGURE 4 Schematic of solar energy system at Riegel Textile Corporation plant in LaFrance, South Carolina.

There are a number of examples of SIPH for textile dyeing. One of those is for a dyeing operation at the Riegel Textile Corporation plant in LaFrance, South Carolina. Figure 4 shows a schematic of the system. Because the dyeing process uses water at a temperature of 70°C, flat plate solar collectors are used in this system. The system has 621 m² of collector area.

The dyeing process at this plant is an atmospheric dye beak batch process at a maximum temperature of 90°C, and this is also typical of the textile industry in the United States. The batch dyeing process involves heating approximately 4500 L of inlet water between 10–25 and 90°C. The typical dye cycle is shown in Table III. The SIPH system at the Riegel Textile plant has been operating successfully. The system was simulated using the TRNSYS (Transient System Simulation) computer model. Figure 5 shows the actual performance of the system and that predicted by TRNSYS. According to measured performance of the system from August 14 to October 8, 1983, the system operated with an efficiency of 46%.

3.2 SIPH System for Milk Processing

The food system uses approximately 17% of the U.S. energy, nearly 50% of which is for food processing as hot water ($<100^{\circ}\text{C}$) or hot air. Proctor and Morse showed that more than 40% of the energy demand in the beverage industry in Australia was in the form of hot water between 60 and 80°C . Considering the temperature and heat requirements of food processing, one would be tempted to conclude that SIPH would be ideal. However, many food processing requirements are seasonal, and this might not be economical considering the prices of conventional fuels unless the SIPH system can be used for a majority of the year. One application that is year-round is milk processing. An example of a TRNSYS simulation of a SIPH system for milk processing in the United States obtained from the literature, with the unit operations (and their respective temperatures) for this plant compatible with solar thermal energy, is as follows:

- Boiler feed makeup water (100°C)
- Pasteurizer makeup water (21°C)

TABLE III

Typical Dye Beck Heat Requirements

Dye process operation	Temperature of dye beck ($^{\circ}\text{C}$)	Time interval of cycle (h)	Percentage of total process energy
Heat the initial load	16–32	0.5	7
Dyeing preparation	32–88	0.5	22
Dye period	88	7–9	68
Cool and reheat for dye fixation	32–43	1.0	3

- Case washer and rinsing (49°C)
- Cleanup (71°C)
- High-temperature short-time cleanup (79°C)
- Bottled water (93°C)

It was estimated that for a plant producing 170,000 kg/week of milk and 98,000 kg/week of orange juice, a total of 621.3 GJ (or 80%) of energy demand was compatible with solar energy during the summer and 724.9 GJ (or 93%) of energy demand was compatible during the winter. A 4000- m^2 flat plate collector area could provide approximately 30 to 35% solar fraction for the milk processing plant in Madison, Wisconsin; Fresno, California; or Charleston, South Carolina.

3.3 SIPH Systems for Electronic Processing

Gould Electronics, a leading U.S. electronic circuit board manufacturer, uses an SIPH system to provide the process hot water requirement in its foil division manufacturing facility in Chandler, Arizona. The SIPH system was built when Arizona had a moratorium on natural gas systems in new buildings. The SIPH system uses PTCs, through which oil is heated and stored in two 30,000-gallon (113,000-L) storage tanks. The collector field has a total aperture of 60,480 square feet (5600 m^2). The plant has a hot water requirement of 265 gallons/min (1000 L/min) at 196°F (91°C). The hot oil is used to heat up this water, with electric induction heaters serving as backup for the times when the solar system is not able to meet the full load. The field operated successfully from 1983 to 1987, but by 1990 many

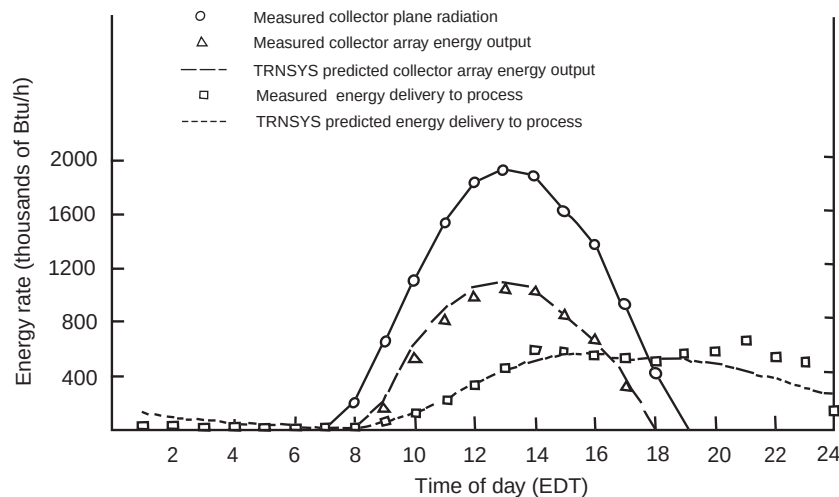


FIGURE 5 Actual and predicted system operation: August 16, 1983.

small problems had resulted in a substantial decrease in field performance. Gould engineers worked with Sandia National Laboratories to solve these problems between 1991 and 1992, restoring system availability to 90%. Based on actual operational data, it was estimated that an average saving of \$7800 per month was achieved by using the solar system.

3.4 Direct Process Heat

Solar energy, with the appropriate choice of collection and concentration technology, can be used to deliver a range of temperatures, from very low to extremely high values. This allows SIPH systems to be designed to provide direct heat to a variety of processes. SIPH systems can be designed to provide temperatures that can exceed values achieved using practical combustion processes.

The use of SIPH has been proposed for various processes, including high-temperature detoxification and recycling of hazardous waste; production of fuels such as hydrogen and syngas through thermochemical cycles; production of metals such as aluminum, iron, zinc, and magnesium; and production of other materials such as silicon nitride and silicon carbide in solar furnaces. The use of concentrated solar radiation has also been proposed for the heat treatment of materials. Although several SIPH systems have been proposed for direct process heat applications, there are no known commercial systems that currently operate using these technologies. Nevertheless, research in this direction has progressed, and several laboratory and pilot plant scale demonstrations of these technologies have been completed.

4. DESIGN AND PERFORMANCE OF SIPH SYSTEMS

To assess the economic viability of any solar process, its cumulative energy delivery over its economic life on the order of years or decades must be known. It is very difficult to calculate this number accurately because (1) solar systems and their energy delivery are subject to the vagaries of local microclimates (which can change on a time scale on the order of hours) and (2) future weather cannot be predicted at this level of detail. The standard approach used to estimate future performance of a solar system is to use a typical year of past weather data and assume that it will represent the future on the average, for

purposes of engineering accuracy. Two common methods used to predict long-term performance of SIPH systems are the utilizability method and TRNSYS. These methods can also be used for the design of SIPH systems. This section describes the utilizability method and the TRNSYS method.

4.1 The Utilizability Method

Utilizability, ϕ , has been used to describe the fraction of solar flux absorbed by a collector that is delivered to the working fluid. On a monthly time scale,

$$\bar{\phi} = Q_u / F \bar{\eta}_o \bar{I}_c < 1.0.$$

Here, the overbar is used to indicate monthly average values, Q_u is the monthly averaged daily total useful energy delivery, F is a heat exchanger transfer factor, $\bar{\eta}_o$ is the daily averaged optical efficiency of the collector, and $\bar{I}_c$ is the monthly averaged daily solar radiation incident on the collectors. $\bar{\phi}$ is the fraction of the absorbed solar flux that is delivered to the fluid in a collector operating at a fixed temperature T_c . The $\bar{\phi}$ concept does not apply to a system composed of collectors, storage, and other components where the value of T_c varies continuously. However, with most concentrators for concentration ratios greater than 10, the collector is relatively insensitive to a small range of operating temperatures. To check this assumption for a particular process, values of $\bar{\phi}$ at the extremes of the expected temperature excursion can be compared.

The value of $\bar{\phi}$ depends on many system and climatic parameters. Collares-Pereira and Rabl showed that only three are of the first order: the clearness index $\bar{K}_T$, the critical intensity ratio $\bar{X}$, and the ratio r_d/r_T . The first is related to insolation statistics, the second to collector parameters and operating conditions, and the third to collector tracking and solar geometry.

Empirical expressions for $\bar{\phi}$ have been developed for several collector types.

For example, for nontracking collectors,

$$\bar{\phi} = \exp\{-[\bar{X} - (0.337 - 1.76\bar{K}_T + 0.55r_d/r_T)\bar{X}^2]\}$$

$$\text{for } \bar{\phi} > 0.4, \bar{K}_T = [0.3 - 0.5], \text{ and } \bar{X} = [0 - 1.2].$$

Here, $\bar{X}$ is the monthly average daily critical intensity ratio, defined as the ratio of collector heat loss to absorbed solar flux at the "no net energy delivery" condition. The monthly averaged clearness index $\bar{K}_T$ is defined as the ratio of the monthly averaged daily total radiation on a horizontal surface to that on a corresponding surface outside the earth's atmosphere.

The ratio r_d/r_T is a factor that depends on solar geometry and collector tracking. Expressions are available and tabulated for various collector types.

Similar equations are available for other collector types and other clearness index values. In some cases, the collection time Δt_c might need to be determined for nontracking, low-concentration collectors.

A well-known method to calculate the long-term performance of collector systems with storage is using the “f-Chart” method developed by Klein and Beckman. The f-Chart method can be used for estimating the fraction of the load that will be met by the solar thermal system. The method uses the utilizability (defined differently from the method given previously) and a few other nondimensional parameters to predict the monthly solar load fraction f_s . Predictions are based on the results of simulations for various climatic conditions by an hourly time scale computer model.

4.2 TRNSYS Computer Simulation Program

TRNSYS is a sequential-modular transient simulation program developed at the Solar Energy Laboratory of the University of Wisconsin. It is a widely used, detailed design tool involving hourly simulations of a solar energy system over an entire year. The TRNSYS computer program contains FORTRAN subroutines of nearly all the components necessary to build a solar energy system. The component models, which are either empirical or analytical, describe the component performance with algebraic and/or differential equations. A system simulation model is created by interconnecting the models of individual components. The resulting set of simultaneous algebraic and/or differential equations is solved by TRNSYS. This software is constantly being upgraded and is backed by technical support.

5. ECONOMICS OF SIPH

The commercialization of SIPH systems did not take off as was expected during the late 1970s and early 1980s. One reason why was the low price of natural gas in the market. Carwile and Hewett analyzed 15 proposed SIPH systems that were never built and the examined reasons why. It was seen that the payback period was too large when compared with industry standards for many of these projects. An additional hurdle that had to be overcome was investors' lack of

confidence in this unconventional method of providing IPH. The small number of existing SIPH systems in the market made it difficult to convince investors to allocate money for the solar systems, even though the economics were favorable. Studies have been performed on the future prospects of SIPH, particularly systems using PTC for collecting solar radiation. As per the analysis, it was forecast that there was considerable potential to reduce the cost of such systems. If costs could be brought down by research and development (R&D) and by larger production volumes, a bigger market share would result in the future. Some studies estimated that with cost reductions through R&D, SIPH could have as much as 35% of the projected market by 2030.

In conclusion, SIPH systems are technically capable of meeting the IPH needs of many of today's industries. However, the higher costs of SIPH systems and the availability of cheaper alternatives have resulted in only few SIPH systems being built and still being in operation.

SEE ALSO THE FOLLOWING ARTICLES

Arid Environments, Impacts of Energy Development in • Materials for Solar Energy • Solar Cells • Solar Cookers • Solar Cooling, Dehumidification, and Air-Conditioning • Solar Detoxification and Disinfection • Solar Distillation and Drying • Solar Heat Pumps • Solar Ponds • Solar Thermal Power Generation • Solar Water Desalination • Thermal Energy Storage

Further Reading

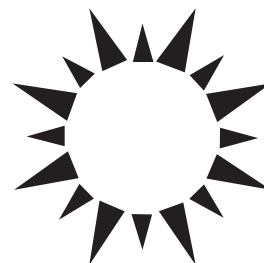
- Carwile, C., and Hewett, R. (1995). “Barriers to Solar Process Heat Projects: Fifteen Highly Promising (but cancelled) Projects,” NREL/TP-472-7295. National Renewable Energy Laboratory, Golden, CO.
- Collares-Pereira, M., and Rabl, A. (1979). Simple procedure for predicting long term average performance of nonconcentrating and of concentrating solar collectors. *Solar Energy* 23, 223–234.
- Duffie, J. A., and Beckman, W. A. (1991). “Solar Engineering of Thermal Processes.” John Wiley, New York.
- Goswami, D. Y., Kreith, F., and Kreider, J. F. (2000). “Principles of Solar Engineering.” 2nd ed. Taylor & Francis, Philadelphia.
- Klein, S. A., and Beckman, W. A. (1979). General design method for closed loop solar energy systems. *Solar Energy* 22, 269–282.
- Kreider, J. F. (1979). “Medium and High Temperature Solar Processes.” 2nd ed. Academic Press, New York.
- Procter, D., and Morse, R. N. (1975). Solar energy for Australian food processing industry. CSIRO Solar Energy Studies, East Melbourne, Victoria, Australia.

Solar Thermal Power Generation

ANDREAS LUZZI and KEITH LOVEGROVE

Australian National University

Canberra, Australia



-
1. The Solar Resource
 2. Main Solar Collector Systems
 3. Energy Collection and Conversion
 4. Developments of Concentrator Systems
 5. Economics
 6. Outlook

Glossary

conduction band Excited electronic status of semiconductor materials, with readiness for electron transport.

heliostats Plane mirrors that continuously adjust in angle according to the sun's position, so as to reflect a beam of solar radiation to some fixed point in space.

kilowatt (kW) A unit of power equal to 1 kJ of energy applied per second. Addition of a subscript "e" indicates electrical energy, subscript "th" indicates thermal energy.

kilowatt-hour (kWh) A unit of energy equal to the power of 1 kW applied over the duration of 1 h. Addition of a subscript "e" indicates electrical energy, subscript "th" indicates thermal energy.

megawatt (MW) A unit of energy equal to 1000 kW.

nanometer A very small unit of length, equal to one one-millionth of a millimeter.

photovoltaics (PV) The science behind semiconductor-based solar energy conversion systems.

pyranometer An instrument that measures the intensity of total solar radiation.

pyrheliometer An instrument that measures the intensity of direct beam solar radiation.

receiver A device that intercepts concentrated solar radiation and transforms it into another form of energy.

secondary concentrator An arrangement of mirrors to form a light funnel to further concentrate preconcentrated solar radiation that is incident on a receiver.

stagnation temperature The equilibrium temperature a solar receiver will reach if the process that normally removes energy is turned off.

thermal loss Undesirable transfer of heat away from a solar receiver to the environment.

valence band Relaxed electronic status of semiconductor materials.

Solar thermal power generation systems capture energy from solar radiation, transform it into heat, and then use an engine cycle to generate electricity. The majority of electricity generated around the world comes from thermally driven steam-based systems. The heat for these systems is largely provided by the combustion of coal, oil, or natural gas or by nuclear fission. The concept of substituting solar radiation for these thermal inputs leads to the consideration of solar thermal power generation. The second law of thermodynamics expresses the empirically observed fact that the potential conversion efficiency from heat to electricity (work) increases with the temperature of the thermal input available. For solar systems, this naturally leads to the consideration of methods for concentrating solar radiation to achieve high temperatures. Small-scale solar thermal power generation systems were demonstrated as early as the 1860s, mainly in France and the United States. In 2003, the largest examples of solar thermal power plants are the 80 megawatt electric (MW_e) parabolic trough systems that have been operating in southern California since the late 1980s. There are nine such power plants with variable sizes that have a combined nominal capacity of 354 MW_e and generate approximately 800,000 MWh (megawatt-hour) of electricity every year. This is sufficient to service a medium-sized city of approximately 200,000 people. These power plants generate electricity on a competitive basis thanks to previous tax incentives, long-term amortization, green-power credits, some natural gas back-up, and expert operation and maintenance.

1. THE SOLAR RESOURCE

The sun is approximately 1.4 million km in diameter and approximately 150 million km from the earth. It is close to 5500°C at its surface and emits radiation

at a rate of 3.8×10^{23} kW. This energy is supplied by nuclear fusion reactions near its core, which are estimated to continue for several billion years.

The major part (70%) of the solar radiation received on the earth's surface is within the visible spectrum, with a maximum intensity at approximately 500 nm. This is consistent with a "black body" radiating at 5500°C. The intensity of solar radiation reaching the surface of the earth is typically between 900 and 1000 Wm^{-2} on a clear sunny day at noon.

Only a tiny fraction (4.5×10^{-8} percent) of the total energy radiated by the sun reaches the outer surface of the earth's atmosphere, but this is still equal to approximately 1.6×10^{14} kW. Putting this in perspective, the total annual world electricity generation of approximately 1.5×10^{13} kWh could be met with a hypothetical solar thermal power generation system (using technology discussed in this article) covering an area of approximately 30,000 km^2 , which is approximately equal to the size of Belgium or Taiwan, for example. The total or "global" solar radiation reaching the earth's surface is made up of "direct" and "diffuse" components. Direct radiation can be concentrated with mirrors or lenses that point directly at the sun. It is measured with a pyrheliometer. Diffuse radiation is that fraction that has been scattered by water vapor, particles, or molecules within the atmosphere. It cannot easily be concentrated and makes up approximately 15% or more of the total radiation, depending on the clarity of the sky. Diffuse radiation levels are measured with a pyranometer fitted with a shading device to exclude direct radiation. Together the diffuse and direct components combine to make up the global total and this can be measured with an unshaded pyranometer.

However, there are challenges associated with turning this solar energy source into continuous electrical power:

- The solar radiation reaching the surface of the earth is diffuse ($\sim 1000 \text{ Wm}^{-2}$ maximum) compared to other sources of energy, such as fossil fuel flames (megawatts/m^2);
- The sun shines intermittently in a daily cycle, which in turn follows a seasonal variation and is also heavily influenced by local weather conditions;
- The position (azimuth and elevation) of the sun in the sky and hence the angle of incidence of direct beam radiation vary continuously;
- Solar radiation cannot be stored directly, but must be immediately converted into heat, electricity, or chemical energy.

Technological solutions that allow these challenges to be met do exist.

2. MAIN SOLAR COLLECTOR SYSTEMS

Various devices for collecting solar radiation thermally have been devised. At the simplest level, a flat metal plate, painted black and placed in the sun, will heat up until it reaches a temperature where the amount of heat that it radiates and loses to the air around it exactly balances the amount of energy it receives from the sun. This "stagnation temperature" occurs at approximately 80°C for a simple flat-plate solar collector. If water, for example, is passed through passages in the plate collector, then it will stabilize at a lower temperature and the water will extract some of the energy in being usefully heated up. This is the essence of solar thermal energy collection.

Greater levels of sophistication are aimed at reducing the amount of "thermal loss" from the collector surface at a given temperature. This allows energy to be collected more efficiently and at higher temperatures. The thermal efficiency of a solar collector can be defined as:

$$\text{Thermal efficiency} = \frac{\text{Energy converted}}{\text{Solar energy intercepted}}.$$

Starting with the flat-plate solar collector, a cover layer of glass helps by cutting down the energy lost by the circulation of cold air across the collector. If metal tubes and glass cylinders are used as collectors instead of plates, then the space can be evacuated, so that air convection losses are largely eliminated. The smart use of coating materials to produce an optically selective surface helps to reduce radiation losses. Such coatings absorb as much as possible of the high-energy (ultraviolet and visible) solar wavelengths while they emit as little thermal radiation (infrared) as possible from the plate. Various combinations of these measures are used in the design of systems for the production of solar hot water that are used around the world for domestic and industrial applications. In principle, solar collectors of this nature could be used for electricity production. However, the comparably low temperatures that are achievable limit the conversion efficiencies that are possible to low levels.

Further increasing the temperature at which energy can usefully be recovered requires some

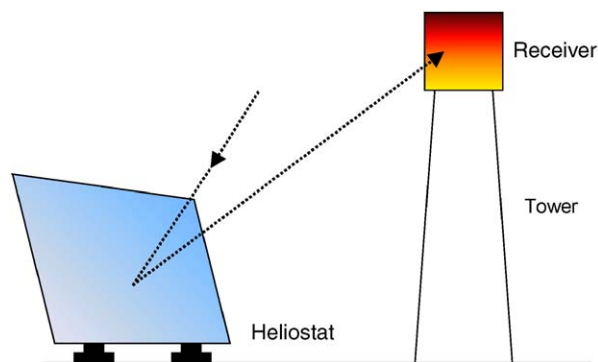


FIGURE 1 The central receiver concept; a field of plane mirror heliostats all move independently to one another to keep a beam of solar radiation focused on a single central receiver.

method of optically concentrating the radiation so that the size of the absorbing surface and hence its thermal loss is reduced. The conceptually simplest approach is to employ a series of flat mirrors, called heliostats, that are continuously adjusted to direct solar radiation onto the absorbing surface. This is illustrated conceptually in Fig. 1. Large plants called “central receiver systems” or “power towers” have been built based on this principle and are discussed in detail in Section 4.1. The concept can also be adapted to linear absorbers and long strips of flat mirrors to create a “linear Fresnel” concentrator.

Alternatively, the mathematical properties of a parabola can be exploited for solar energy concentration. The equation for a parabola in the x - y plane is:

$$y = x^2/4f.$$

Rays of light parallel to the y axis of a mirrored parabola will all be reflected and focused at the focal point at a distance f from the vertex. As illustrated in Fig. 2, this effect can be used in a linear arrangement, where a mirrored “trough” with a parabolic cross section will concentrate solar radiation onto a line-focus when it is pointed directly at the sun. The largest solar thermal power plants thus far constructed employ this principle. They are discussed in Section 4.2. Alternatively, a mirrored dish with a parabolic cross section (a paraboloid) will focus solar radiation to a point focus (see Fig. 3). Paraboloidal dish systems are discussed in Section 4.3. Both dishes and troughs require continuous adjustment of position (or at least frequent readjustment) to maintain the focus as the sun moves through the sky.

Similar focusing effects can obviously be achieved with lenses of various kinds, but this has not been employed on the scales needed for solar thermal power systems.

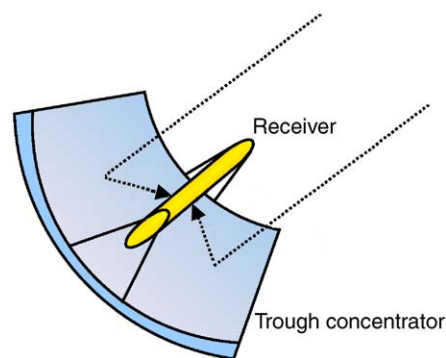


FIGURE 2 A parabolic trough concentrator focuses solar radiation onto a linear receiver when faced directly toward the sun.

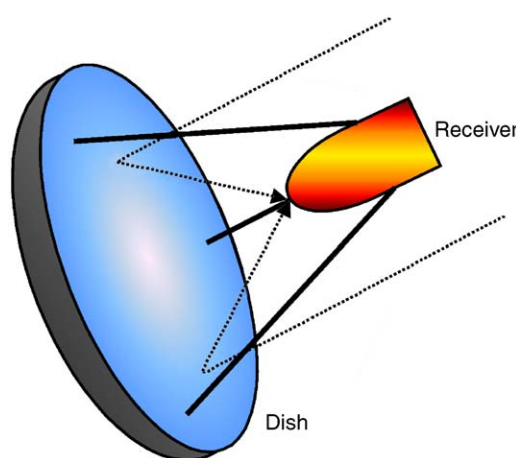


FIGURE 3 A paraboloidal dish concentrator focuses solar radiation onto a point focus receiver.



FIGURE 4 A nonimaging concentrator concentrates solar radiation without the need to track the sun.

Another alternative, which potentially avoids the need to track the sun, is to employ “nonimaging” concentration. As illustrated in Fig. 4, this involves the construction of a mirrored “light funnel” of some

TABLE I

Typical Operating Temperature and Concentration Range of the Various Solar Thermal Collector Technologies

Solar collector technology	Typical operating temperature (°C)	Concentration ratio	Tracking	Maximum conversion efficiency (Carnot)
Flat-plate collector	30–100	1	—	21%
Evacuated-tube collector	90–160	1	—	38%
Solar pond	70–90	1	—	19%
Solar chimney	20–80	1	—	17%
Linear Fresnel reflector technology	260–400	8–80	One-axis	56%
Parabolic trough collectors	260–400	8–80	One-axis	56%
Heliostat field + central receiver	500–800	600–1000	Two-axes	73%
Paraboloidal dish concentrators	500–1200	800–8000	Two-axes	80%

kind. Such a device will be able to collect rays into its aperture over a range of incidence angles and cause them to exit through a smaller aperture via multiple reflections. Nonimaging concentrators have thus far not found application as the “primary” means of concentration for solar thermal power systems but they are frequently applied as “secondary concentrators” at the focus of central receivers, dishes or troughs, where they serve to further reduce the size of the focal region.

The rays of light from the sun are not exactly parallel. This means that even a perfect optical system will produce an image of finite size, with an intensity distribution that is a maximum in the center and tapers off to zero at the edges. Imperfections in mirror shape and tracking accuracy have the effect of further spreading out the sun’s image.

Each of these approaches to solar energy collection has a typical ratio of collected radiation intensity to incident solar radiation intensity, termed the “concentration ratio.” Table I summarizes the options discussed and lists typical concentration ratios, the resultant operating temperatures, and the consequent thermodynamic limiting efficiency with which electricity could be produced. The limiting conversion efficiency arises from the second law of thermodynamics. The maximum efficiency for conversion of heat from a constant high temperature source is given by:

$$\text{Maximum conversion efficiency} = 1 - T_{\text{cold}}/T_{\text{hot}}.$$

This is the “Carnot limit.” (Note that “absolute” temperatures must be used in the equation, usually degrees Kelvin, equal to °C + 273.15.) Real solar thermal power generation systems typically achieve approximately one-third or less of the ideal maximum Carnot efficiency.

Although higher concentration ratios give higher efficiency, they also lead to potentially higher complexity and cost. The ultimate challenge with solar thermal power systems is to produce the desired output as economically as possible. This invariably means that a tradeoff between system efficiency and capital investment drives the design process.

The three main solar thermal concentrating technologies are discussed in detail in this article as they constitute the bulk of the commercial development efforts undertaken in the area of solar thermal power generation.

3. ENERGY COLLECTION AND CONVERSION

A solar thermal power system can be presented schematically as shown in Fig. 5.

All systems begin with a concentrator, followed by a receiver that catches the concentrated radiation and converts it to another form (typically thermal energy). At this point there are two options: either the energy is further converted to the final form desired (i.e., electricity) or it is transported to another location for final conversion. The choice of the transport path provides the option of storage in the intermediate form before going to final conversion. There is also the option of designing an energy storage system after final conversion.

3.1 Receiver Types

Receivers are often built with a cavity geometry; this gives a higher effective “absorptivity,” maximizing the amount of radiation absorbed and also reducing the surface area susceptible to radiation and

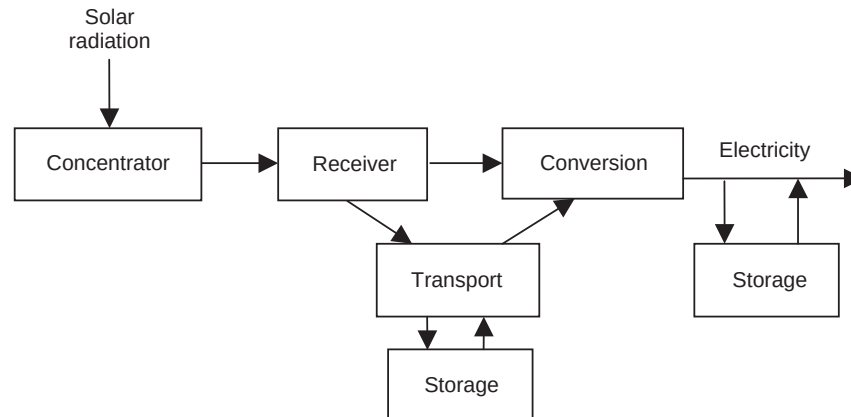


FIGURE 5 Schematic representation of the component parts of a solar thermal power system.

convection losses. Notable exceptions include the single-tube receivers with evacuated insulation used on the Californian parabolic trough power plants and the cylindrical externally irradiated receivers used on power tower systems with a completely surrounding heliostat field. There are several methods for construction. Four of the commonly used approaches are discussed here.

3.1.1 Tubular Receivers

The conceptually simplest thermal receiver type involves circulating the working fluid or heat transfer fluid through directly irradiated tubes. The tubes can be arranged to form a cavity by winding a single tube to form a helical cylinder. An alternative is arranging a number of parallel straight tubes around the outside.

Tube receivers provide a simple and reliable way of designing for high system pressures. There are issues associated with thermal cycling (particularly for steam receivers) and flow balancing in multiple parallel tube receivers.

3.1.2 Volumetric Receivers

Volumetric or direct absorption receivers aim to have the radiation absorbed by surfaces directly in contact with the working fluid. This is often achieved by having a window in front of a cavity receiver that contains an absorbing matrix, over which a fluid passes. Windows are generally made of quartz because of its superior ability to handle thermal gradients. Matrices are often ceramic (such as a foam or ceramic formed into spikes). Such receivers have been used for chemical reactors, with a catalyst impregnated in ceramic foam, and for heating air. If air is being used at atmospheric pressure, then the window is not essential.

3.1.3 Heat-Pipe Receivers

A major issue with receivers is matching the incident flux to the requirements of the energy conversion system. A very effective approach to this is the “heat pipe” receiver. The idea is to use a container with a liquid/vapor mixture. The container will be essentially all at the boiling temperature corresponding to the pressure. In regions where heat is added, liquid will evaporate, whereas it will condense where heat is extracted. The result is a “black box” of extremely high effective conductivity. Sodium (boiling point 883°C at atmospheric pressure) is favored for solar applications. This approach has been used with central receiver systems, Stirling engines, and chemical reactor receivers.

3.1.4 Solid-State Receivers

Solid-state receivers are based on the application of semiconductor materials that are designed as directly irradiated photovoltaic or thermoelectric devices. These systems produce low-voltage direct current electric power that is then transformed to higher voltage alternative current for use.

3.2 Heat Transport Mechanisms

Various methods are employed for the task of energy transport within solar thermal power plants. Five major heat transfer mechanisms for solar application are discussed here.

3.2.1 Oil

The Californian parabolic trough power plants employ a synthetic heat transfer oil. The oil is circulated through the receivers and brought back to the centrally located power plant where a heat exchanger is used to raise steam for power

generation. It has the advantage of providing stable and predictable receiver operation. Oils, however, have the disadvantages of slowly decomposing when operated at high temperatures and of environmental as well as fire hazards when leaking. The oil lines must be adequately insulated and design allowance must be made for adequate pipe expansion.

3.2.2 Steam

If the final conversion process is based on a steam Rankine cycle, then it makes sense to collect the energy as steam in the first place. The major challenge with direct steam generation, however, is the rapid thermal cycling that can occur in tubular receivers. At the point in the receiver where the tubes “dry out,” the internal convective heat transfer coefficient is much less than in the regions where tubes are in contact with water. Problems arise because some sections of the tubing are exposed to rapidly alternating exposure to vapor and liquid. This causes localized thermal cycling and leads to fatigue of the metal. Insulation of pipe work and allowance for expansion is needed, as are sophisticated rotary joints to allow for solar concentrator tracking.

3.2.3 Air

Central receiver systems have been designed, among other options, to use air as heat transfer fluid either in pressurized tubular systems or in atmospherically open volumetric receivers. Systems with air temperatures of up to 1000°C have been extensively tested. Whereas tubular receivers allow for high-pressure operation with associated potential for conversion efficiency gains in gas turbines, the volumetric air receiver concept achieves very effective heat transfer from sunlight to air with minimized thermal gradients. Metallic mesh receivers have been used for temperatures of up to approximately 700°C and ceramic matrix elements for higher temperatures. Hot air from volumetric receivers can be used to heat chemical reactor tubes or boiler tubes in exactly the same way as if they were fired with fossil fuels. This provides for a uniform and controllable method for heating.

3.2.4 Molten Salt

A direct competitor to the air-based system for central receivers is molten salt. It offers the advantages of a high-specific-heat, single-phase fluid in the receiver and can also provide effective energy storage using insulated tanks. A disadvantage is that all pipes must have a provision for external heating to keep

the system warm overnight and to facilitate cold starts (melting point of typical nitrate salts vary between approximately 140 and 260°C). Molten salts are corrosive and react with air and water if leaks occur. The usual nitrate salts employed, though, are environmentally benign fertilizer salts.

3.2.5 Chemical

Thermochemical cycles with fluid reactants can provide energy transport and storage. If energy is converted chemically at the receiver, then the use of counterflow heat exchangers means that reactants to and from the receiver are at ambient temperature. Hence, chemical systems not only provide an elegant way of storing energy, they also provide simple heat-loss-free energy transport. Pipes no longer need insulating, they neither expand nor contract, and much simpler mechanisms such as flexible joints are able to get through the axes of rotation.

3.3 Conversion Systems

There are three main thermo-mechanical conversion systems that are being applied with solar thermal power technologies. These are Rankine cycle, Stirling engine, and Brayton cycle systems.

3.3.1 Rankine Cycle

Steam-based Rankine cycles are responsible for the majority of electric power generation in the world. The technology is readily applicable to solar thermal systems as long as the energy collected can be transported to a central power block. Water is first compressed by a feed-water pump and then boiled and superheated (up to approximately 500°C) before being expanded through a turbine that turns an electric generator. The low-pressure steam is then condensed in a heat exchanger and fed back to the feed-water pump to be reused. Steam Rankine cycles have been, and continue to be, employed with parabolic trough, paraboloidal dish and central receiver solar thermal power plants.

Organic Rankine cycles are identical to steam-based Rankine cycles with the exception that organic fluids are employed. Organic fluids such as toluene have a lower liquid-to-gas phase change temperature than water/steam and therefore employ a lower-temperature heat input of typically between 80 and 280°C. Flat-plate, trough, dish, and solar pond systems have all been demonstrated to be suitable solar collector systems for organic Rankine cycle power generation.

3.3.2 Stirling Engines

The most extensive development of engine systems for direct solar operation has been based on the Stirling cycle, invented in 1816. The Stirling cycle employs external heating and cooling of its working fluid. The ideal cycle is made up of a constant-volume heating step from a sink temperature to a source temperature, augmented by isothermal heating, followed by a constant-volume cooling step to a sink temperature, and finished by further isothermal cooling. Dish/Stirling systems achieve very high net solar-to-electric conversion efficiencies, with the solar power world record for a precommercial 25 kilowatt electric (kW_e) unit being 29.4%.

In solar applications, Stirling engines commonly use helium or hydrogen as the working fluid. Kinematic (crankshaft-coupled) and “free-piston” engine versions are the two main concept variants developed worldwide. In a free piston engine, a floating-drive piston bounces back and forth between gas volumes at either end of a single cylinder. The piston itself serves as a linear generator. The result is a minimum number of moving parts. Solar receiver designs for Stirling engines have employed either directly irradiated fine-tubing for heating the working fluid or heat transfer heating via sodium heat pipe systems. Hybridization with fossil-fuel firing for 24 h operation has been demonstrated.

3.3.3 Brayton Cycle

The Brayton cycle is the basis of the conventional gas turbine. The cycle involves adiabatic (i.e., insulated) compression of a gas by a compressor turbine, constant pressure addition of heat, and adiabatic expansion in an expansion turbine (usually attached to the same shaft as the compressor), followed by constant pressure cooling. In a fuel-fired system, the heat addition is carried out in a combustion chamber and gases are exhausted after expansion, either with or without heat recuperation. For solar applications, heat recuperation is economically warranted for efficiency gains and gas fuel back-up advised for system control purposes as well as on-demand operation regardless of time of day.

3.3.4 Others

There are many other ways to convert thermal energy into work, many of which can potentially be applied to solar thermal power generation systems. They include, for example, the following:

- Thermoelectric converters produce electricity directly from heat. Semiconductor-based systems

work in an analogous way to photovoltaic (PV) cells. Excitation of electrons from the valence band into the conduction band of the semiconductor material occurs via thermal excitation rather than individual photon absorption;

- Thermo-photovoltaics use PV cells tailored to infrared thermal radiation wavelengths to convert the radiation re-emitted from heated surfaces; and
- Magneto-hydrodynamic converters use the expansion of heated, ionized gas through a magnetic field to generate a potential difference.

4. DEVELOPMENTS OF CONCENTRATOR SYSTEMS

The three major solar thermal concentrating technologies, central receivers, parabolic troughs, and paraboloidal dishes, are discussed in more detail in this section.

4.1 Central Receiver Systems

The central receiver (or power tower) concept was first proposed by scientists in the U.S.S.R. in the mid-1950s. The first experiment was established in Sant’ Ilario near Genoa, Italy, in 1965 by Professor Giovanni Francia. He installed 120 round mirrors, each the size of a “tea table,” focusing on a small steam generator on top of a steel frame. The product was superheated steam (500°C, 10 MPa).

Central receivers have the advantages that large-area radiation collection to a central point occurs optically and that all the energy conversion takes place at this single fixed point. This avoids the need for energy transport networks and allows investment to improve the efficiency and sophistication of the energy conversion process to be made more cost-effectively. Associated disadvantages are that they must be built as single large systems, without the modularity benefits of distributed systems, such as troughs or dishes. The fixed position of the receiver also means that heliostats do not point directly at the sun for the majority of the time. The amount of collected solar radiation per unit area of mirror is therefore less than with the other solar concentrating technologies.

Major investigations during the past 20 years have focused on four heat transfer fluid systems. They are water/steam, sodium, molten salt, and air (see Section 3.2). An overview of the main grid-connected central receiver research facilities and demonstration

TABLE II

Summary of Grid-Connected Central Receiver Research Facilities and Demonstration Power Plants Built over the Past 20 years

	Net electric power	Total reflector area	Heat transfer fluid	Effective storage capacity (electrical or thermal)	Period of service
Eurelios (Italy)	1 MW	6216 m ²	Water/steam	0.5 MWh _e	1980–1984 ^a
Sunshine (Japan)	1 MW	12,912 m ²	Water/steam	3 MWh _e	1981–1984 ^a
IEA-CRS (Spain)	0.5 MW	3655 m ²	Sodium	1.0 MWh _e	1981–1985 ^b
Solar One (United States)	10 MW	71,447 m ²	Water/steam	28 MWh _e	1982–1988 ^c
Solar Two (United States)	10 MW	81,344 m ²	Molten salt	107 MWh _{th}	1996–1999 ^d
CESA-1 (Spain)	1.2 MW	11,880 m ²	Water/steam	3.5 MWh _e	1983–1984 ^b
Themis (France)	2.5 MW	10,794 m ²	Molten salt	12.5 MWh _e	1983–1986 ^b
MSEE (United States)	0.75 MW	7845 m ²	Molten salt	2.5 MWh _e	1984–1985 ^b
SES-5 (former Soviet Union)	5 MW	40,584 m ²	Water/steam	1.5 MWh _e	1985–1988 ^b
Weizmann (Israel)	0.5 MW (since 2001)	3472 m ²	Beam-down concept	None	1988–today ^b

^a Dismantled.^b Continued operation as research facility until today.^c Converted to Solar Two.^d Mothballed.FIGURE 6 A 120 m² heliostat for central receiver power plants. Source: Sanlucar-120.

power plants built over the past 20 years is given in Table II.

Thus far, power generation has been via conventional steam Rankine cycles at the base of the tower. In early systems, tubular receivers were built to produce superheated steam directly. Thermal problems associated with the unsteady boundary between liquid water and steam in the tubes, however, motivated a move to secondary heat transfer fluids, such as sodium, molten salt, or air. Liquid sodium provides good heat transfer behavior, but carries the disadvantage that all the transport piping must be heated above sodium's melting point (close to 100°C). The possibility of leaks also poses a significant fire risk. The use of molten salt avoids

the fire risk but is well suited to storage in tanks to allow power generation when there is no sun. As a downside, the nitrate salt mixtures commonly used have a melting point of between 140 and 260°C and corrosion challenges arise if leaks occur. Being the most environmentally friendly heat transfer fluid, air has been extensively and successfully tested at scales of up to 3 megawatt thermal (MW_{th}).

Heliostat fields can either surround the tower or be spread out on the shadow side of the tower. Whereas externally irradiated cylindrical receivers are predominantly employed for surround fields, cavity receivers are used with one-sided heliostat fields. System designers have developed optimization strategies that determine the best arrangement for a given number of heliostats. These take into account the effects of shading between heliostats during the course of the day, the spread of the field, and the optical inefficiency that increases as heliostats are further from the tower.

Two generic approaches to heliostat design have been used. The most obvious is a plane structure with rigid mirror facets mounted on it. The structure sits on top of a pedestal with a drive arrangement that allows for two-axes tracking of the sun. The other alternative is termed the "stretched-membrane" approach. As the name implies, membranes are stretched across a circular frame in a manner similar to a drum skin. Thin stainless steel sheets covered with thin-glass mirror tiles are used successfully. A variety of reflective polymer films have also been tested over the years, but face issues of limited

durability. Focusing of this heliostat concept is achieved by applying a small vacuum inside the drum. In this way heliostats can focus their own sun image onto the tower receiver rather than directing a plane beam at it. Figure 6 illustrates an example of a rigid mirror facet heliostat design. Development trends have suggested that larger heliostats might be more cost-effective. Current optima are between 100 and 150 m².

The largest central receiver solar thermal power plant demonstrated thus far is the “Solar Two” plant in southern California. This plant is an updated version of the previously operated “Solar One” system. Extra heliostats were added and the receiver was converted from direct steam generation to molten salt. Figure 7 shows the Solar Two plant in operation and Fig. 8 is a schematic illustrating the

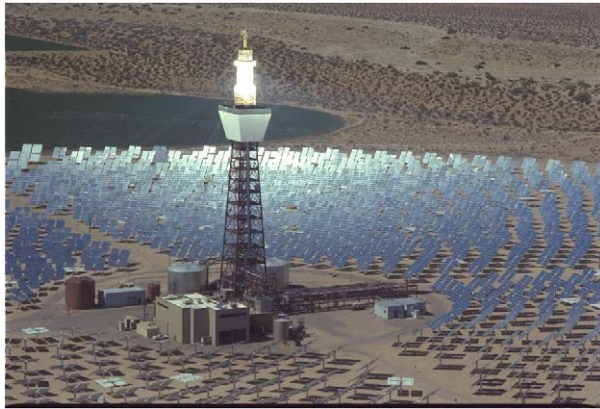


FIGURE 7 The “Solar Two” (previously “Solar One”) central receiver power plant in operation. Source: Sandia National Laboratories.

operating principles. When the solar field is operating, the molten nitrate salt moves from a “cold” (288°C) storage tank via the receiver at the top of the tower, where it is heated, to the hot storage tank (566°C). Independently of the solar energy collection process, salt from the hot tank is passed through a heat exchanger where heat is transferred to produce superheated steam, with the salt passing back to the cold storage tank. The steam is used in a conventional steam turbine power plant for electricity generation.

The Solar Two plant has 1018 heliostats of 39.1 m² plus a further 108 heliostats of 95 m². Under nominal conditions, 48 MW_{th} is concentrated onto the receiver that sits at the top of a 91 m high tower. Steam is produced in the heat exchangers at 10 MPa and 538°C and the net electrical output is 10.4 MW_e.

One of the newest developments in the central receiver area is the “beam-down” concept proposed and tested in part by the Weizmann Institute of Science in Israel. Rather than converting the concentrated solar energy at the top of the tower in a receiver, a hyperbolically shaped secondary mirror directs the converging radiation vertically downward to a focal point at the bottom of the tower. On the ground, an additional nonimaging mirror concentrates the radiation further before the concentrated sunlight is captured by a volumetric receiver. This receiver is capable of reaching very high air temperatures to power, for instance, a Brayton cycle gas turbine.

4.2 Trough Systems

Solar thermal power in the form of mechanical energy for water pumping was established for the

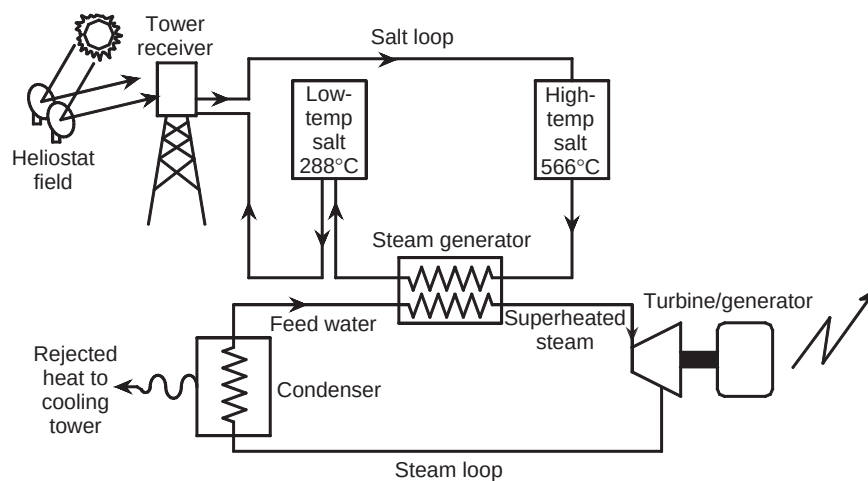


FIGURE 8 Schematic of “Solar Two” operation.

first time by Frank Shuman near Cairo in 1913 (~ 40 kW). It incorporated a water/steam operated parabolic trough array (five units, $4\text{ m} \times 62\text{ m}$ each) and a low-pressure condensing steam engine. Solar electric trough development was continued in the mid-1970s by the U.S. Department of Energy. The first experimental system started operation in 1979. At the same time, a private research and development company from Israel decided to design and commercialize parabolic trough “solar electric generating systems” (SEGS) in southern California. This decision was strongly motivated by favorable power purchase agreements and tax credits offered in the state of California.

Nine SEGS power plants were built between 1984 and 1989, with a combined capacity of 354 MW_e. These investor-owned, natural gas-assisted power plants have not only been operating on a fully commercial basis ever since, but have also been continuously improved. They operate beyond their nameplate capacity, generating important solar peaking power for the southern California grid, at the world’s most competitive solar power prices.

During the early 1980s, small, trough-based solar thermal demonstration power systems were con-

structed in the United States, Japan, Spain, and Australia. Table III lists the details of these plants. Specifications for the nine SEGS plants are given in Table IV.

Figure 9 shows one of the SEGS plants and Fig. 10 illustrates the operating principles schematically. A synthetic heat transfer oil is pumped through the trough solar collector array and heated to 400°C . This oil is then used to produce steam in heat exchangers before being circulated back to the array. The steam is used in a conventional steam turbine-based electricity-generating plant. Although some hot oil-based energy storage was provided in the first plant, the SEGS systems overall rely on natural gas firing to provide continuous operation when the sun is not available. The special tax law for these power plants limits the fossil-fuel cofiring to a maximum of 25% per annum of the total heat energy supplied to the boilers.

The SEGS troughs are built using a galvanized steel space frame. This frame supports 4 mm thick toughened glass mirror facets that are shaped by heating and molding the raw glass to match the parabolic profile, before silvering. Each mirror facet is supported at four attachment points. The most

TABLE III

Details of the Main Small, Trough-Based Solar Thermal Demonstration Power Plants

	Net electric power	Total aperture area	Heat transfer fluid	Effective storage capacity (electrical or thermal)	Duration of service
Coolidge (United States)	0.15 Mw	2140 m ²	Synthetic oil	5 MW _{th}	1980–1982
Sunshine (Japan)	1 Mw	12,856 m ²	Water/steam	3 MW _e	1981–1984
IEA-DCS (Spain)	0.5 Mw	7622 m ²	Synthetic oil	0.8 MW _e	1981–1985
STEP-100 (Australia)	0.1 Mw	920 m ²	Synthetic oil	117 MW _{th}	1982–1985

TABLE IV

Details of the Californian Parabolic Trough “SEGS” Solar Thermal Power Plants

	Net electric power	Total aperture area	Duration of service
SEGS I	13.8 MW	83,000 m ²	1984–present
SEGS II	30 MW	19,000 m ²	1985–present
SEGS III	30 MW	230,000 m ²	1986–present
SEGS IV	30 MW	230,000 m ²	1986–present
SEGS V	30 MW	251,000 m ²	1987–present
SEGS VI	30 MW	188,000 m ²	1988–present
SEGS VII	30 MW	194,000 m ²	1988–present
SEGS VIII	80 MW	464,000 m ²	1989–present
SEGS IX	80 MW	484,000 m ²	1989–present

recently constructed systems (termed LS-3) are 5.76 m wide and 95 m long, resulting in a total aperture of 545 m² each. Two hundred twenty-four glass mirror facets are used and a geometric concentration ratio of up to 80:1 is achieved. The trough units are lined up in north-south rows and track the sun from east to west during the course of the day. Each parabolic trough has its own positioning and local control system.



FIGURE 9 View of SEGS trough-based solar thermal power plant in southern California. Source: Kramer Junction Operating Company.

The receiver units of the SEGS troughs consist of a stainless steel tube 70 mm in diameter, covered by a Pyrex glass envelope that is sealed to the tube via metal bellows at each end. The space between the steel and glass is evacuated to minimize thermal losses. The surface of the stainless steel absorber tubes is coated with a selective surface that absorbs up to 97% of the incident solar radiation while minimizing the amount that is radiated at thermal (infrared) wavelengths. The combination of trough and receiver is capable of operating at temperatures in excess of 400°C. However, the synthetic heat transfer oil becomes chemically unstable and begins to break down at temperatures above 300°C. Approximately 350 m³ of oil circulates in each of the 30 MW_e plants. Depending on the operating regime, this oil may need to be replaced at rates of up to 6% per year as a result of thermochemical breakdown.

Each of the latest SEGS plants employs a steam turbine with a reheat cycle and multiple extractions. With steam inlet conditions of 10 MPa and 370°C, thermal-to-electric conversion efficiencies of approximately 37% are achieved, giving overall peak solar-to-electric conversion efficiencies of up to 24%.

Regular maintenance is required for all solar thermal power plant systems. For the SEGS plants, routine cleaning and replacement of broken mirror facets and receiver modules forms a major part of the maintenance program. Major improvements have been realized over time.

Research and development on parabolic trough systems has continued since the SEGS plants were completed. A major area of investigation has targeted the replacement of the heat transfer oil with direct

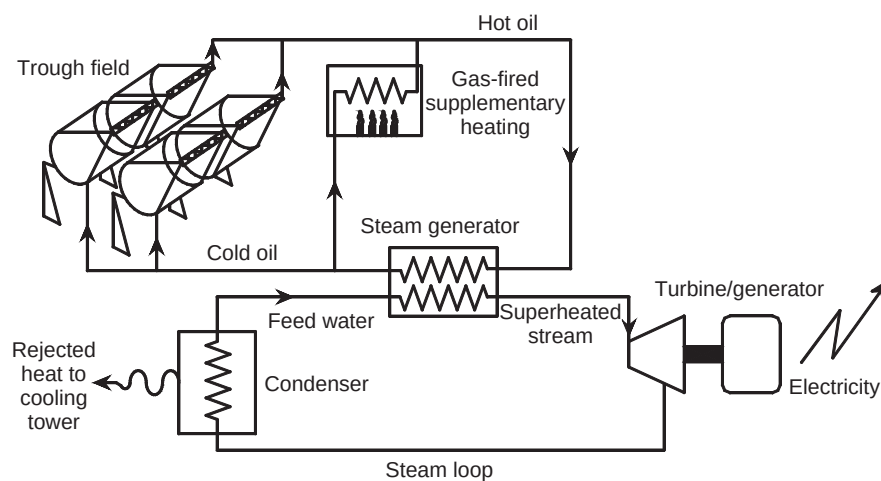


FIGURE 10 Schematic representation of SEGS plant operation.

generation of steam in the receivers. Direct steam generation allows collection of energy at higher temperatures as well as elimination of one heat-exchange step and thus improves the efficiency of the steam turbine. It also avoids the need to replace the costly oil and eliminate the inherent fire risk. When attempting once-through-to-superheat boiling, the challenge is in managing the rapidly changing thermal stresses induced in the receiver tube by the unsteady, localized movement of the liquid/vapor phase boundary when boiling is occurring.

In a variation of linear focusing technology, compact Fresnel-type concentrator systems are being developed at various places around the world. This concept dates back to the 1950s and is much like a linear version of a central receiver system. Fixed linear receivers are illuminated by a series of long, narrow mirrors that track the sun individually in a single axis. If several receiver rows are installed side by side, then the individual mirror units can switch from one receiver to another depending on the relative optical efficiencies at different times of the day. Alternatively, a secondary concentration of the reflected light near the receiver can be used to ease precision and costs of the linear primary solar concentrator.

4.3 Paraboloidal Dishes

The first realization of a solar dish concentrator/engine system design dates back to 1864. The Frenchman Augustin Mouchot built a variety of “thermo-piles” and “solar engines” using conical mirrors. There are many designs and states of development of paraboloidal dishes.

Paraboloidal dish concentrators must point directly at the sun for proper operation. This necessitates two-axes tracking mechanisms. There are two main approaches to this: “altitude/azimuth” tracking or “polar/equatorial” tracking. In the first approach, the dish is mounted with a horizontal axis pivot (to allow altitude adjustment) plus a vertical axis pivot (to allow azimuth adjustment). A polar/equatorial system has a main axis of rotation, the angle of which is adjusted on a daily basis so that it is at a right angle to the sun’s noon elevation. With the polar axis correctly adjusted, tracking during the day involves following the sun from sunrise to sunset on a constant basis with a single axis of movement. Altitude/azimuth arrangements have the disadvantage that both actuators must work throughout the day and the azimuth adjustment must be relatively rapid near midday for high sun elevations. Despite

this, it is sometimes used for pragmatic structural design reasons.

Dish-based solar thermal power systems can be divided into two groups: those that generate electricity with engines at the focus of each dish and those that use some mechanism to transport heat from an array of dishes to a single central power-generating block. From a research and development perspective, dish systems have two major advantages over central receivers in particular, in that a single dish prototype can be constructed and tested relatively economically. Furthermore, for the dish/engine approach, a single module on its own can demonstrate the operation of the full system. Consequently, there have been well over 40 different dish prototypes developed thus far, by both research institutes and private engineering companies.

The variations in structural design of dish concentrators are very similar to the different types of heliostats used with central receivers. Possible different approaches include the following:

- A space frame structure that supports a number of rigid mirror elements held in a paraboloidal orientation;
- A single rigid paraboloidal surface covered with mirror elements;
- A single, mirrored, stretched-membrane that is pulled into a near-paraboloidal shape on a drum-like structure; and
- A space frame supporting a series of small stretched membrane mirror facet elements.

The high concentration ratios achievable with dish concentrators allow for efficient operation at high temperatures. Stirling cycle engines are well suited to construction at the size needed for operation on single-dish systems and they function with good efficiency, with receiver operating temperatures in the range of 650 to 800°C. To achieve good power-to-weight ratios, working gas pressures in the range 5–20 MPa are employed and the use of high-conductivity gases, such as hydrogen or helium, gives improved heat transfer. Dish/Stirling units of 25 kW_e have achieved solar-to-electric conversion efficiencies of close to 30%. This represents the maximum net solar-to-electric conversion efficiency yet achieved by any nonlaboratory solar energy conversion technology.

Figure 11 shows two dish/Stirling system operating in southern Spain. The dish is fitted with a 10 kW_e two-cylinder kinematic Stirling engine. One-millimeter-thick, back-silvered glass mirror facets are employed to achieve very high focusing accuracy.



FIGURE 11 Paraboloidal dish solar concentrator with a receiver-mounted Stirling engine operating in southern Spain. Source: Schlaich, Bergemann, and Partner.

Including various alternative versions of the generic system, over 80,000 h of cumulative on-sun operating time have thus far been clocked with dish/Stirling systems.

The largest distributed array/central plant solar thermal power system that has yet been demonstrated is the “Solarplant-1” system built in southern California in 1983–1984. A photograph of this system is shown in Fig. 12. It consisted of 700 dishes with a total collecting area of 30,590 m². The dishes generated steam in their cavity receivers, with 600 of the dishes producing saturated steam at approximately 6 MPa (275°C) and the remainder taking the saturated steam to a superheat of 460°C. The steam was transported through an insulated pipe network to a central steam turbine-based generating plant that produced a nominal output of 4.9 MW_e. The dishes were constructed with multiple stretched-membrane mirror elements. The plant ceased operation in 1990.

The world’s largest solar dish, with an aperture of 400 m², is shown in Fig. 13. It was developed for use in distributed systems similar to Solarplant-1. Two prototypes are in operation: one at the Australian National University in Canberra, Australia and the other at the Sde Boquer Desert Research Centre of the Ben Gurion University in Israel. These dishes have triangular mirror panels supported on a hexagonal aperture space frame structure. Altitude–azimuth tracking is employed, with the horizontal axis near the base of the dish so that it can be parked in a horizontal position relatively close to the ground. This helps to reduce wind resistance and thus improve storm survivability. The Canberra dish has a cavity receiver based on a single helical winding of tubing to serve as a once-through-to-superheat boiler producing superheated steam at 5 MPa and



FIGURE 12 Seven hundred dishes producing steam for a central turbine system. Source: LaJet.



FIGURE 13 The world’s largest 400 m² solar dish system. Source: Australian National University.

500°C. Although intended for operation in large arrays, the prototype system is connected to a small reciprocating steam engine that is capable of generating up to 45 kW_e for the local electricity grid. The thermal efficiency from sun to steam is between 85 and 90% and, in a large steam turbine-based system, production of electricity at a rate of approximately 100 kW_e per dish would be expected.

Research and development on paraboloidal dish systems around the world strive to improve and develop dish designs to reduce the costs and to optimize the reliability and efficiency of conversion systems. A number of groups are investigating the use of small solar-driven gas turbines at the focus of dishes (dish/Brayton systems). Such systems offer the potential for high-efficiency operation and moderate maintenance requirements.

Chemical conversion systems are also investigated using paraboloidal dish concentrators. High-temperature reversible chemical reactions, such as

methane re-formation (up to 900°C) or ammonia dissociation (up to 700°C), can be used to store heat energy in chemical form. The resulting chemical products, such as hydrogen and carbon dioxide, are stored and transported at ambient temperature and thereafter recombined in an exothermic reaction to provide the heat for power generation on a continuous, 24 h basis.

5. ECONOMICS

Solar thermal power generation technologies have achieved impressive development progress over the past two decades. Some of this progress, most notably with parabolic troughs, has resulted in successful technology commercialization, delivering the lowest-cost solar power yet achieved.

Given the very many different technology routes, development status, application options, and project sizes, solar thermal power generation technologies cannot be compared, let alone summarized, on a one-line economic basis. It is, however, valuable to look at the achievements, status, and trends of capital costs, operation and maintenance costs, and levelized electricity costs. These costs are strongly dependent on the country and on site-specific labor-rate structures. Although some general observations can be made, it is not necessarily valid to translate costs experienced with a particular project in one country to operation in another by a simple exchange-rate conversion.

5.1 Capital Costs

Capital costs arise for land, site preparation, construction work, solar collectors, power conversion units, the balance of plant including controls, and infrastructure as well as services such as engineering, project management, and commissioning.

Depending on the plant capacity, its technical complexity, the chosen location (geophysics, climate, access, water, etc.), and the choice of storage and/or fuel hybridization or back-up, solar thermal power plant capital costs over the past two decades have varied greatly between approximately U.S. \$3000 and U.S. \$9000 per kilowatt of nominal plant capacity.

Economies-of-scale progress with the Californian SEGS parabolic trough plants has shown that a cost reduction of typically between 12 and 16% has been achieved on doubling of the plant capacity.

In 2003, 10 to 50 MW_e power plants are offered for approximately U.S. \$2500–\$3000 per kilowatt,

whereas 0.5 MW_e to 10 MW_e power plants require investment capital of U.S. \$3000–\$5000 per kilowatt of nominal plant capacity. Smaller projects, such as one-off 10–25 kW_e dish/Stirling units, cost approximately U.S. \$6000–\$7000 per kilowatt.

Given that solar thermal power generation technologies are comparably “low-tech” offerings based primarily on structural and civil engineering skills and efforts, cost reduction curves are quite predictable. Performance improvement, manufacturing progress, economies-of-scale, and options of hybridization as well as energy storage dominate cost reductions, indicating trends toward capital costs of approximately U.S. \$1500–\$2500 per kilowatt.

5.2 Operation and Maintenance Costs

Operation and maintenance (O&M) costs are composed of labor and materials costs. The split between these two varies according to technology choice, remoteness of location, labor rates, and water costs, with typical ratios of approximately 1:2 for large power plants and 2:1 for small power plants.

Whereas current commercial solar thermal power plants have typical O&M costs in the range of U.S. \$0.025–\$0.035 per net kilowatt-hour of power generated, continuous optimization efforts indicate a strong trend toward U.S. \$0.01–\$0.015 per kilowatt-hour for utility-scale power stations in the range of 50–100 MW_e over the next decade.

5.3 Levelized Electricity Costs

Levelized electricity costs (LECs) are averaged electricity costs calculated on a simple basis of fixed economic parameters and a straight-line depreciation of capital cost over the duration of the project life. Technologies as well as projects are difficult to compare due to the large variety of design- and finance-specific parameters chosen.

Nevertheless, when analyzing well-demonstrated, megawatt-scale projects on a comparative basis of 30 years of project life, 8% real discount rate, 1% insurance rate, straight-line depreciation, and 0% income tax rate, typical LECs of U.S. \$0.12–\$0.16 per kilowatt-hour result. Small-scale, one-off projects, such as dish/Stirling systems, are likely to achieve LECs of U.S. \$0.20–\$0.30 per kilowatt-hour once technically matured and commercially operated. Following the ongoing efforts and predictions of major project developers, megawatt-scale solar thermal power plants are likely to ultimately generate LECs of U.S. \$0.05–0.1 per kilowatt-hour.

6. OUTLOOK

The future of solar thermal power generation holds great commercial promise and a broad spectrum of interesting new opportunities. Among these possibilities are 24 h solar-only power generation thanks to thermal and/or thermochemical energy storage, the re-formation and synthesis of high-value fuels and fine chemicals, the thermal and/or thermochemical production of hydrogen, and the detoxification of hazardous waste.

SEE ALSO THE FOLLOWING ARTICLES

Arid Environments, Impacts of Energy Development in • Cogeneration • Hybrid Energy Systems • Materials for Solar Energy • Solar Cells • Solar Cooling, Dehumidification, and Air-Conditioning •

Solar Energy, History of • Solar Fuels and Materials • Solar Heat Pumps • Solar Ponds • Solar Thermal Energy, Industrial Heat Applications • Solar Water Desalination • Thermal Energy Storage

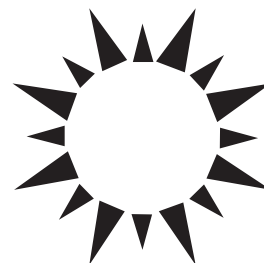
Further Reading

- Becker, M., Gupta, B., Meinecke, W., and Bohn, M. (1995). "Solar Energy Concentrating Systems, Applications, and Technologies." C. F. Mueller Verlag, Heidelberg, Germany.
- Gordon, J. (ed.). (2001). "Solar Energy, the State of the Art, ISES Position Papers." James and James, London.
- International Energy Agency's Solar PACES (Solar Power and Chemical Energy Systems) Web site: www.solarpaces.org.
- Meinel, A. B., and Meinel, M. P. (1976). "Applied Solar Energy: An Introduction." Addison-Wesley, Reading, MA.
- Pilkington Solar International. (1996). "Status Report on Solar Thermal Power Plants." Pilkington Solar International, Cologne, Germany.
- Winter, C.-J., Sizmann, R. L., and Vant-Hull, L. L. (eds.). (1991). "Solar Power Plants: Fundamentals, Technology, Systems, Economics." Springer Verlag, New York.

Solar Water Desalination

E. DELYANNIS and V. BELESSIOTIS

National Center for Scientific Research "Demokritos"
Athens, Greece



-
1. Historical Background
 2. Solar Stills
 3. Principles of Solar Stills
 4. Heat and Mass Balance in the Still
 5. Prediction of Daily Output
 6. Materials of Construction
 7. Multistage Solar Stills
 8. Solar-Driven Desalination
 9. Energy Conversion Systems
 10. Conversion to Electricity
 11. Sizing a Solar-Driven System
 12. Conclusion

Glossary

electrodialysis Desalination method using membranes to separate anions and cations from water by applying an electric field to the membranes.

evacuated tube collectors Medium-temperature solar radiation collectors (up to 250°C) using selective absorption surfaces into the vacuum tubes.

flat-plate collectors Devices that absorb solar radiation on black surfaces and transfer the thermal energy to a fluid. They are low-temperature devices up to 95°C.

focusing collectors Devices that focus solar radiation on a surface or a point. The high-temperature thermal energy is transmitted to pressurized water or it produces vapor.

photovoltaic Devices that transform solar radiation into electricity by the use of semiconductors.

renewable energies Inexhaustible natural energy resources (solar, wind, geothermal, tidal, wave, and biomass), replaceable as they are consumed. These exist in huge amounts worldwide.

reverse osmosis Pressure-driven desalination method using semipermeable membranes that permit only water to be transferred through the membrane, and the solutes are retained.

solar collectors Devices that absorb solar radiation in select black surfaces and transform it to thermal energy, which is transferred to a suitable fluid.

solar ponds Small artificial lakes with salt gradient water that absorb solar radiation and act as storage devices for hot water and as devices supplying hot water with a temperatures up to 90°C.

Solar energy desalination is a method by which the sun's energy is used to desalinate brackish or seawater to produce fresh drinkable water. There are two methods for using solar energy: directly by heating and evaporating the brackish or seawater in a solar still (this method is called solar distillation) and indirectly by capturing solar energy using one of the techniques that transform solar radiation into thermal or electrical energy to drive a conventional desalination method (the indirect method is called solar-assisted or solar-driven desalination). Indirect thermal solar desalination methods are the multistage-flash thermal distillation method, the multiple-effect thermal distillation method, and the thermal vapor compression method. Electrical power generated from the sun's energy is used to drive reverse osmosis, electrodialysis, or mechanical vapor compression desalination methods. Theoretically, solar energy can replace any other energy source, but from a practical standpoint this is not yet technically feasible, especially for large-capacity solar desalination plants. This is due to the spreading and very low concentration of solar radiation falling on the earth's surface and variations in solar incident radiation. Solar energy depends on local insolation rates, which vary depending on location, the hour of the day, the day of the month, and the month of the year. These conditions result in an unsteady state diurnal operation. On the other hand, because solar radiation is very diluted, large surface areas are needed to capture the necessary amount of energy. Storage of heat or electricity could be a solution for smooth 24-h operation, but it adds to the capital cost, which in many cases is prohibiting. Thus, solar desalination installations are of small- or medium-size capacity, ranging from a few to 100 m³/day, and in some cases up to 300 m³/day (26,420–79,260 gallons per day) for

indirect methods. They can supply with fresh water remote, arid, or semiarid regions that have high intensity solar radiation.

1. HISTORICAL BACKGROUND

Solar-driven desalination was developed mainly in the second half of the 20th century, although solar conversion systems have been known since antiquity. It was developed mainly during the energy crisis in the 1970s. Intensive studies led to the construction of compact, small- and medium-capacity conventional desalination plants that could be driven by solar energy. However, solar distillation has been known since antiquity but only as theory or as a very primitive application. Aristotle described the “water cycle” in nature, which is an immense solar distillation procedure, concluding that this procedure may be used to produce sweet water from salt water. He explained, “as the vapors formed by the evaporation from the surface of seas and lakes rise to the atmosphere and condense again, they give sweet water.” Later, there were some practical applications and references about the production of drinkable water from sea water using solar radiation. The first practical application of solar distillation was the American patent issued in the 19th century to N. M. Wheeler and W. W. Evans. The patent describes a solar still and is a complete scientific source about production of fresh water by distillation using solar energy. No significant changes have been made to this theory.

The first large solar distillation plant was built in Las Salinas, Chile, to provide fresh water from mine brackish water with 140 g/kg total dissolved solids (TDS) (140,000 ppm). The plant had a surface area of 4450 m² (48,000 ft²), produced 22.7 m³/day [$\sim$ 6000 gallons per day (GPD)] of fresh water for the mine and the people working in it, and was in operation for approximately 40 years until the mine was closed.

Until the early 1960s, no other large-scale application of solar distillation was reported. Nevertheless, during World War II distillation gained interest due to the lack of fresh water for troops of Africa and the islands of the Pacific region. In 1945, at the Massachusetts Institute of Technology (MIT), Maria Telkes invented individual inflatable solar stills that saved many lives. After World War II, the MIT team developed various solar still configurations, which have been studied at the office of saline water (OSW) demonstration plant in Daytona Beach, Florida. Some commercial installations have been

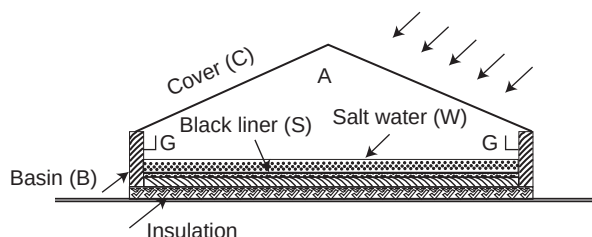


FIGURE 1 The greenhouse symmetrical-type solar still.

erected in St. John, Virgin Islands, and Australia. In the 1960s, the Church World Service erected two distillation plants in the Greek Islands, where four additional solar distillation plants have been built that produce fresh water for the local communities..

2. SOLAR STILLS

Since the 1980s, no large-capacity solar distillation plants have been built, mainly for the following reasons: Large installation areas are needed, solar distillation plants have low efficiencies, and solar-driven desalination has been developed, which in some cases can replace solar distillation (e.g., for medium-size communities).

A solar still is composed of simple equipment, which acts simultaneously as a converter of solar radiation into thermal energy and provides storage for the heat. There are two types of solar stills: single- and multiple-effect stills. For both types, a large variety of configurations have been studied, but most have no practical applications. Studies have focused mainly on increasing output by various means, but this leads to an increase in capital or operation costs. Thus, simple, single-effect stills are the only ones used for large-scale applications. They can be easily constructed from local materials and easily operated without skilled personnel.

There are two types of single-effect solar stills (called greenhouse solar stills): symmetrical (Fig. 1) and asymmetrical (Fig. 4). These types have many configurations depending on the geometry of the still and the type of cover (glass or plastic). Figure 2 shows some of the single-effect solar still designs used to construct solar distillation plants.

3. PRINCIPLES OF SOLAR STILLS

A solar still consists of the following:

A well-insulated basin (Fig. 1B), which may be made from metal, plastic, treated wood, concrete, or

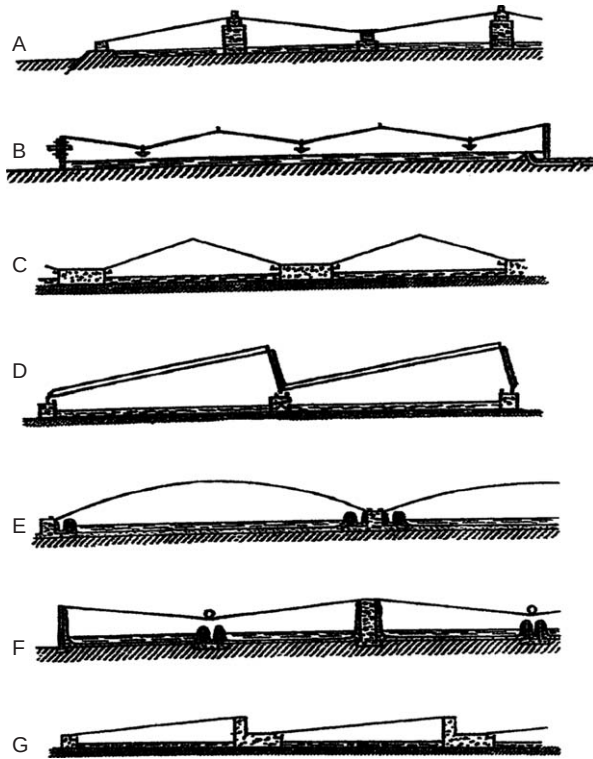


FIGURE 2 Various configurations of single-effect solar stills used to construct solar distillation plants: (A) Battele-Loef glass-covered type; (B) Howe University glass-covered type; (C) CSIRO Australian glass-covered type; (D) Technical University, Athens, glass-covered type; (E) Inflated, tedlar plastic film type; (F) V-shaped plastic cover design; and (G) stretched-inclined plastic cover type.

any material that withstands temperatures of 90–100°C, according to the local availability.

A black sheet (Fig. 1S) that covers the inside of the bottom of the basin to avoid leakage of hot water. The basin can be deep or shallow and contains the water to be heated by solar radiation. The black liner absorbs the incident radiation.

A frame that supports the transparent cover. The frame can be made of metal (e.g., aluminum), treated wood, concrete, or plastic material.

A tilted transparent cover (Fig. 1C) made from either glass or a plastic sheet that withstands ultraviolet radiation and is not easily deteriorated by heat.

Collecting canals, the gathers (Fig. 1G), placed inside the still walls to collect the condensate of fresh water, running along the transparent cover.

Various auxiliaries, such as tubes for loading the still with salt water and for unloading to drain the brine, water storage tanks, and distilled water pumps.

Solar radiation penetrates through the tilted transparent cover into the airtight space (Fig. 1A)

and is partially absorbed by the salt water in the basin, which is heated to a temperature higher than that of the cover but lower than boiling point of salt water. The air-vapor mixture leaving the saltwater surface has a higher temperature and lower density than the air-vapor mixture directly beneath the cover. Thus, a convection current is formed and the saturated air-vapor mixture from the water surface moves upward, is cooled by contact with the colder cover, becomes saturated, and is partially condensed, forming a thin film that slides along the inside cover surface by gravity. The uncondensed air-vapor mixture moves back to the water surface. The process takes place in narrow layers between the two surfaces, but the bulk of the air mass does not participate because diffusion and heat conductance are low. The mechanism of the vapor flow in the space takes place by convection and the optimum space is the one that enhances conduction. Therefore, it is advantageous to keep the distance between the water surface and the cover, as well as the angle between the cover and horizontal, as small as possible to achieve good convection (normally, the lowest angle is 15°). It is obvious that the productivity of the still depends on the proper sliding of the condensate on the inner surface of the cover and the temperature difference between the saltwater surface and the cover. The larger the difference, the more intense the circulation of the air-vapor mixture becomes and consequently the vapor condensation. Thus, for a greenhouse-type solar still, it is essential to understand the convection phenomena inside the still. Flow patterns and isotherms are given in Fig. 3 for a single solar still with a 15° angle of inclination and an Ra number of 10^{10} .

4. HEAT AND MASS BALANCE IN THE STILL

Mathematical calculations and models for unsteady state conditions in a solar still are very complicated. There are numerous relations and models for heat and mass transfer coefficients to predict the output of a still—almost as many as still designs. Most of them are empirical or semiempirical and are applied only to the specific design of the still. For the single, greenhouse type, the most common is that by Dunkle. Under the assumption that the temperature of the transparent cover T_c and that of the surroundings T_a are uniform, the thermal network in the basin describing the fluxes in a

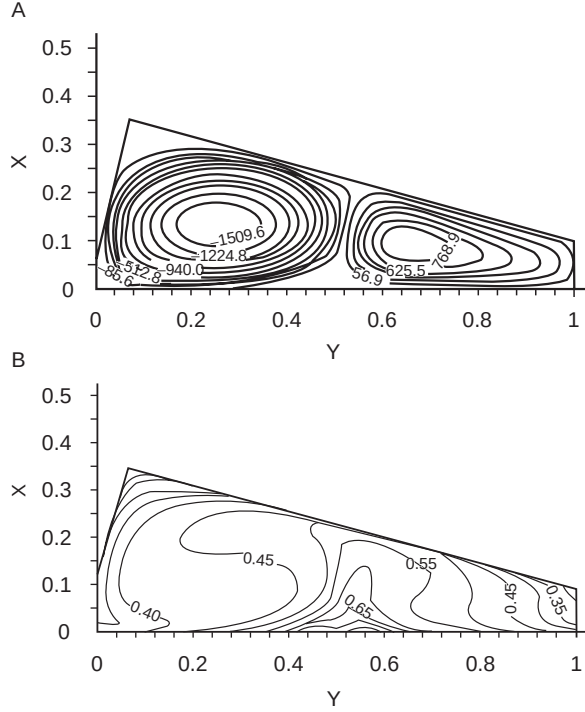


FIGURE 3 Computed streamline (A) and isotherms (B) for an asymmetrical single solar still.

greenhouse-type solar still is presented in Fig. 4. According to the figure, the energy balance for a surface area of the cover A_c and solar radiation intensity I , is

$$\text{Around the cover : } A_g I + q_{rwg} + q_{ewg} = q_{cgs} + q_{rgs} \quad (1)$$

Around the bottom of the still :

$$I \tau_g A_w = q_{rwg} + q_{ewg} + q_b + \frac{C_w \times \partial T_w}{\partial t}, \quad (2)$$

where the last term is the thermal lag for unsteady state operation and is of significance only if the water in the basin is deep. Heat transfer coefficients for radiative heat transfer h_{rgs} , for convective heat transfer h_{cwg} , and for evaporative heat transfer h_{ewg} from the water surface to the cover, as well as radiative heat transfer from the cover to the sky h_{rgs} , for the previous equations are given (in $\text{W/m}^2\text{K}$) as follows:

For radiation :

$$h_{rgw} = 0.9 \cdot \sigma (T_w^2 + T_g^2) \cdot (T_w + T_g) \quad (3)$$

For convection :

$$h_{cwg} = \left[(T_w - T_g) + \left(\frac{p_w - p_g}{268.9 \times 10^3 - p_w} \right) \cdot T_w \right]^{1/3} \quad (4)$$

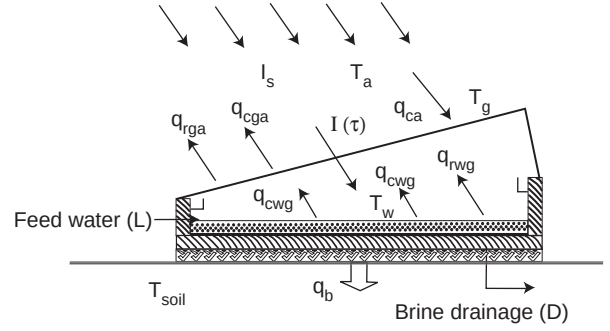


FIGURE 4 Energy balance in a solar still of the asymmetrical type.

For evaporation :

$$h_{cwg} = \frac{9.15 \times 10^{-7} \cdot h_{cwg} \cdot (p_w - p_g) \cdot \Delta h_v}{(T_w - T_g)} \quad (5)$$

The radiative heat transfer between the cover (T_g) and the sky (T_s), which is intensive at night, is given as

$$h_{rgs} = \epsilon_g \cdot \sigma \cdot (T_g^2 + T_s^2) \cdot (T_g + T_s), \quad (6)$$

where Δh_v is the enthalpy of evaporation, and the sky temperature T_s is given as $T_s = 0.0552 \times T_a^{1.5}$. The overall heat transfer coefficient for the upward heat flow for a solar still is given as

$$U_t = [(1/U_i) + (1/A_r \cdot U_o)]^{-1}, \quad (7)$$

where U_i is the overall heat transfer coefficient between the water surface and the cover given as

$$U_i = h_{rgw} + h_{cwg} + h_{ewg} \quad (8)$$

U_o is the overall heat transfer coefficient between the water surface and the environment given as

$$U_o = [h_w \cdot (T_g - T_a) / (T_g - T_s)] + h_{rgs} \quad (9)$$

and the surface area A_r is the ratio of cover area A_g to the free horizontal surface area of the salt water in the basin A_w :

$$A_r = A_g / A_w. \quad (10)$$

The heat balance within the solar still is thus given by the following equation:

$$\frac{C_w}{A_w} \cdot \frac{dT_w}{dt} = \theta(t) \cdot \eta_0 \cdot I(t) - U_t [T_w(t) - T_a(t)] - U_b [T_w(t) - T_a(t)] \quad (11)$$

5. PREDICTION OF DAILY OUTPUT

Operating efficiencies of solar stills in general are low and do not exceed 35%. This is due mainly to the low incident solar radiation per square meter and to the rejection to the atmosphere of the condensation

enthalpy. Consequently, single-effect solar stills have daily output ranging from approximately 2 in wintertime to 5 or 6 liters/m²/day in summertime. Double- or multiple-effect stills have efficiencies up to 46%. Many attempts have been made to increase the output of the stills, including

- Coupling the stills to solar collectors to increase feed water temperature.
- Use of a double glass cover (Fig. 5) with running water inside the cover to increase the temperature difference between the water surface and the cover.
- Trickle feeding of salt water by using wicks or using multiple-effect stills to win condensation enthalpy.
- Use of storage incorporated to the still or in connection with solar collectors to extend operation during nighttime and/or cloudy days.
- In places where rainfall occurs during winter-time, rain may be collected by outside canals. Thus, the output of distilled water may be doubled.

Nevertheless, the output by these means is not always sufficiently high to offset additional capital and operating expenses, and the single-effect, greenhouse-type solar stills remain the most simple and low cost for large-scale applications.

The daily output of a solar still is a function of solar radiation intensity, the duration of daily radiation, and heat fluxes within the still, which in turn depend on the design characteristics of the still. Solar radiation for a bright, cloudless day is given by the curve of daily incident radiation. It can be

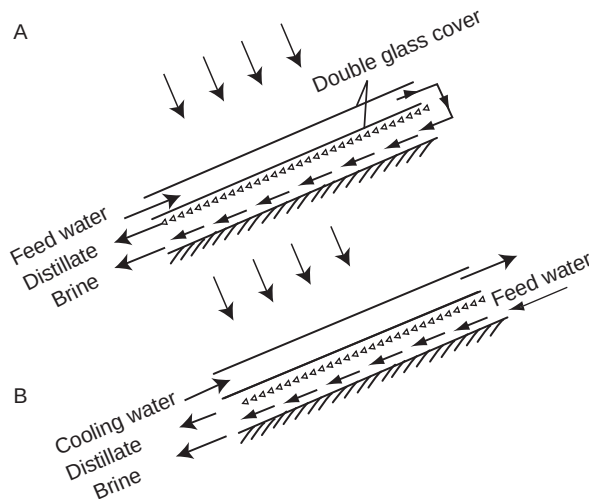


FIGURE 5 Double glass cover solar stills using (A) feedback flow and (B) counter flow.

divided into three phases: phase 1, the early morning start of operation; phase 2, the main operation period called the pseudo-steady state; and phase 3, the saturation of the still at the end of the day during sunset. The mean daily output can be calculated from the following equation, which may be applied to many still configurations if parameters f_1 – f_3 are known:

$$M_{m-d} = f_{1-d} \cdot H_d + f_{2-d} \cdot (\bar{T}_{w-d} - \bar{T}_{a-d}) + f_{3-d}, \quad (12)$$

where H_d is the daily solar radiation (MJ/day); T_w and T_a are the daily mean temperatures of salt water and the environment, respectively; and f_1 – f_3 are characteristic parameters of the still. f_3 concerns the energy state of the still during phase 1 and represents the inertia of the still. Parameters f are functions of the characteristics of the still and of the heat fluxes. They are given as

$$f_1 = \bar{\theta} \cdot n_0 \cdot \frac{A_g \cdot A_w}{C_w} \cdot \frac{h_{ewg}}{\Delta h_v} \cdot \frac{U_t}{U_i} \cdot \Delta t \quad (13)$$

$$f_2 = (U_t - U_b) \cdot \frac{A_g \cdot A_w}{C_w} \cdot \frac{h_{ewg}}{\Delta h_v} \cdot \frac{U_t}{U_i} \cdot (\Delta t)^2 \quad (14)$$

$$f_3 = A_g \cdot \frac{h_{ewg}}{\Delta h_v} \cdot \frac{U_t}{U_i} \cdot (\Delta t)_I \cdot (T_{w-in} - T_{a-in}), \quad (15)$$

where θ is the mean radiation incident inclination angle.

6. MATERIALS OF CONSTRUCTION

To achieve long life and good operation conditions, the construction materials are important. Good maintenance is also important. Thus, the material used must be of good quality, low cost, and, if possible, available in local markets. Small poor communities cannot afford the high capital cost of special materials and long-distance transport expenses. Local inexpensive materials were used to construct solar distillation plants in the Caribbean Islands and India. Thus, from an economic standpoint, solar distillation plants are not comparable from place to place because cost varies according to the materials used. Nevertheless, there are some basic requirements for achieving a long life of operation and the best possible daily output:

- The parts of the still must be precut and easily assembled at the site.
- Maintenance and proper operation must be supervised continuously to achieve an effective life-time of the plant of approximately 15–20 years.

- The system must withstand severe weather conditions, such as hail, hurricanes, and sand and wind storms. Thus, it has to be well anchored in the soil or pavement.
- Construction materials that come in contact with fresh or salt water must not be toxic and in general not contaminate the water produced.
- The still parts must be easily bound in compact packages in case transportation to the site is necessary.

The parts of the still must have the following characteristics:

1. Transparent cover, usually made from glass and, less often, from plastic sheet material:

- The material must withstand wind, sunlight, rain, dust, hail, and/or the wind. Plastic material may be easily deteriorated by sunlight or damaged by hail or wind.
- It must have high transmissibility (more than 90%), low absorptivity, and low reflectivity.
- The properties of the material must not alter with time and exposure to sunlight and environmental effects and must withstand temperatures up to 90°C.
- Especially plastic material should not have electrostatic properties. Electrostatic properties facilitate the accumulation of dust particles, thus decreasing solar radiation absorption inside the still.

2. Solar still frame and basin material, which in most cases are made from the same material. Concrete, wood impregnated with a protective coating, hard plastic material, or any metallic material (e.g., aluminum-magnesium alloy) that is not corroded by distilled or seawater may be used.

- The material must withstand corrosion of distilled and seawater, inside and outside of the still.
- The frames must be anchored, especially in places with strong winds.
- They have to be easily removed or permit access for cleaning and maintenance.

3. The liner of the basin: The black sheet liner acts as an absorber surface for the solar radiation. A variety of materials have been used, including black paints resistant to seawater, butyl rubber sheets, and jute impregnated with tar. In general, the liner must have the following properties:

- It must have absorptivity >0.95 and be impervious to water.
- The sheet surface must be smooth. Rough surfaces facilitate scale deposition and are more difficult to clean.
- It must not deteriorate by solar radiation and must not release toxic substances.

Some auxiliary materials are used, such as pumps, piping systems for drainage and for freshwater transportation, freshwater reservoirs, and sealants. These materials are available commercially. They must be resistant to fresh and seawater and must be cleaned often to avoid algae formation. Formation of algae is dangerous because it results in decreased efficiency and may contaminate fresh water. In stagnant conditions, even inside the still algae are deposited.

7. MULTISTAGE SOLAR STILLS

Recovery and utilization of condensation enthalpy are achieved in multiple-effect solar distillation systems. There are a variety of configurations for double- and multiple-effect solar stills. None have found practical application in large-scale solar distillation plants. Figure 6 demonstrates a two-effect still, and Fig. 7 shows the multieffect wick. A special type is the multieffect diffusion still. In this type, the gap between the cover and saltwater surface is reduced and evaporation occurs only by molecular

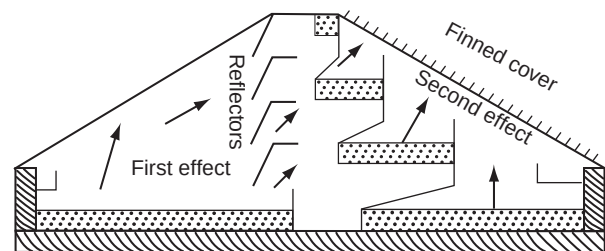


FIGURE 6 A two-effect solar desalination still.

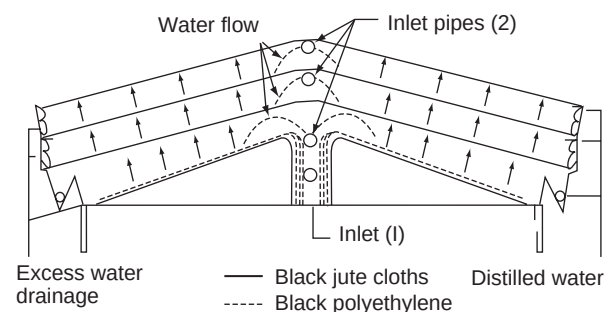


FIGURE 7 Multiple-effect wick solar still.

diffusion and not by natural convection. To determine whether to use a single- or a multieffect still, one must calculate the increase in output and the additional costs in construction and operation.

8. SOLAR-DRIVEN DESALINATION

For medium-size communities, water demand may be for households, agriculture needs, and small industries. These regions may also be remote without energy supply. In this case the solution is solar-driven desalination plant. Capacities range from 10 to 100 m³/day (2642–26,420 GPD).

Solar-driven desalination plants combine two different technologies: thermal or electrical techniques and conventional desalination methods. Desalination methods are considered mature and commercially available, with a variety of procedures, equipment, and capacities. Solar thermal and electrical transformation methods are more or less mature and in some cases commercially available. In comparison to solar distillation plants, both are capital- and operation-intensive enterprises, but costs decrease as technology improves. In general, they are more complicated to operate than plants driven by conventional energy sources. This is due mainly to the unsteady-state solar incident radiation.

9. ENERGY CONVERSION SYSTEMS

Energy conversion systems must be chosen according to the morphology of the site, the desalination method, the availability of salty water, and the freshwater output needed. Wind energy may be included as an energy transformation source because it is an indirect solar energy source. Energy conversion systems include

- Flat-plate collectors, which supply hot water up to ~95°C.
- Solar ponds, which produce hot water up to 90°C.
- Evacuated tube collectors, which reach temperatures up to 200°C and are more suitable for conventional distillation plants, such as multiple-effect thermal distillation (MED) and thermal vapor compression (TVC). They have been proven to be very effective in combination with MED.
- Focusing collectors, which produce high-temperature vapor and/or electricity. They are suitable for large-scale dual-desalination plants and for electricity and freshwater production. Thus, small remote regions are excluded.

Flat-plate collectors and solar ponds require large installation areas (either flat or cascading for flat-plate collectors), but if available, capital and operation costs are low. Not all solar conversion systems are suitable for each desalination method. The combined technologies have to be carefully designed to give the highest possible efficiency for the morphology and environment of the site. Flat-plate collectors, solar ponds, and evacuated tubes are coupled with MED, TVC distillation equipment, focusing collectors, and evacuated tubes for multi-stage-flash thermal distillation plants.

10. CONVERSION TO ELECTRICITY

Solar energy conversion to electricity includes wind energy, which is an indirect solar energy source. The conversion systems comprise photovoltaic devices that directly transform solar energy into electricity and wind turbines that indirectly produce electricity. The electrical energy produced is used to operate seawater and/or brackish water reverse osmosis units, electrodialysis and electrodialysis reversal units, and mechanical vapor compression distillation units. Electrical energy is also used to run auxiliary equipment, such as pumps, fans, and control devices, and to provide the site with electricity. In the case of thermal conversion systems, small photovoltaic installation may help provide the necessary electricity to the system.

Solar-driven desalination plants are unsteady-state systems. Solar energy can be stored as: sensible heat in water tanks, in phase change material, and in batteries as electrical energy to smooth diurnal operation during night and/or cloudy days.

11. SIZING A SOLAR-DRIVEN SYSTEM

For remote and arid locations, solar energy and desalination technology systems have to be as simple as possible, controlled and maintained by skilled personnel. This facilitates good operation conditions and minimal maintenance. Generally, an overcapacity design is recommended for desalination plants and freshwater storage tanks. For small- and medium-size plants, a water storage tank three to six times the daily water consumption and an overcapacity design of approximately 50% are sufficient. Conventional desalination plants are energy-intensive enterprises.

TABLE I

Indicative Typical Energy Inputs for Solar Desalination^a

Process	PR (kg/2326 kJ)	kJ/kg of fresh water	Production rate (liters/m ²)	Form of energy	Suitable energy technology
Stills	0.53	4259	4	50–80°C	Ponds, flat-plate
MED	8	294–296	8	~71°C	Ponds, flat-plate
	15	159–161	13	~120°C	Evacuation-concentrating
MSF	8	298–300	9.5	~90°C	Ponds, evacuated
	12	199–201	14	~120°C	Concentrating
MVC	58	61	50	17 kWh m ⁻³	Solar electric, PV
SWRO	54	43	100	12 kWh m ⁻³	PV, wind
BWRO	200	12	75	3 kWh m ⁻³	PV, wind
BWED	200	12	30	3 kWh m ⁻³	PV, wind

^a Abbreviations used: MED, multiple-effect distillation; MSF, multistage flash distillation; MVC, mechanical vapor compression distillation; SWRO, sea water reverse osmosis; BWRO, brackish water reverse osmosis; BWED, brackish water electrodialysis; PV, photovoltaic.

TABLE II

Solar Drive Desalination Plants Worldwide^a

Plant location	Desalination process	m ³ /day	Solar energy conversion plant
Abu-Dhabi, UAE	MEB, 18 effects	120	Evacuated tubes
Al-Ain, UAE	MED, 55 effects, MSF, 75 stages	500	Parabolic trough
Borkum Island, Germany	MVC	5–12.5 m ³ /h	Wind, 300 kW
Cabo Verde	MSF, auto flash	300	Solar pond
El Paso, Texas	MSF	19	Solar pond
Gran Canaria, Spain	MSF	10	
Hzag area, Tunisia		0.1–0.35	Flat-plate collectors
La Desire Island, French Caribbean	MED, 14 effects	40	Evacuated tube
Lampedusa Island, Italy	RO	16	100 kWp
Margerita di Savoya, Italy	MSF	50–60	Solar ponds
Platforma Solar, Almeria, Spain	MED, 14 effects	72	Parabolic trough
Safat, Kuwait	MSF	10	Solar collectors
Takami Island, Japan	MED, 16 effects	16	Flat-plate collectors
University of Ancona, Italy	MED	30	Solar pond

^a Abbreviations used: MEB, multiple-effect boiling; MED, multiple-effect distillation; MSF, multistage flash distillation; MVC, mechanical vapor compression distillation; RO, reverse osmosis.

Thus, energy needs are the most important factor in determining the size of a solar conversion system. For multieffect stack (MES) evaporators, which are the most preferable to combine with high-performance flat-plate and evacuated tube collectors, thermal energy requirements depend on performance and capacity:

$$Q_{ev} = M_d \cdot \frac{\Delta h_v}{R}, \quad (16)$$

where M_d is the daily rated capacity, and R is the performance ratio of the MES unit, which depends on the number of effects and is given as

$$R = -0.809 + 0.93N - 0.0091N^2. \quad (17)$$

In addition, electrical energy is provided by either diesel or photovoltaic arrays to drive the pumps. Table I gives prices of energy

consumption for solar distillation and solar-driven systems.

12. CONCLUSION

Solar distillation systems are suitable for remote, arid, or semiarid regions for small capacities. Simple designs are preferred because they are cost-effective and easier to operate and maintain. Low-cost, locally available materials can be used, taking into consideration safety and lifetime of the installation. Although some solar distillation plants with capacities of 1–40 m³/day (in Greece, India, the Caribbean islands, and other places) are in operation, the interest currently has turned to solar-driven systems. Solar-driven desalination is suitable for small- to medium-size capacities. As these systems combine, different technologies must be carefully selected and coupled to achieve high efficiencies, simple operation, and easy maintenance. Most plants are pilot plants used to study and improve design, operation, and cost. They do not exceed 0.1% of the capacity of desalination plants driven by conventional energy sources and require further study to improve technologies and lower cost. Table II lists some of the most important solar-driven desalination plants.

SEE ALSO THE FOLLOWING ARTICLES

Desalination and Energy Use • Solar Cooling, Dehumidification, and Air-Conditioning • Solar Detoxification and Disinfection • Solar Distillation and Drying • Solar Thermal Energy, Industrial Heat Applications • Solar Thermal Power Generation

Further Reading

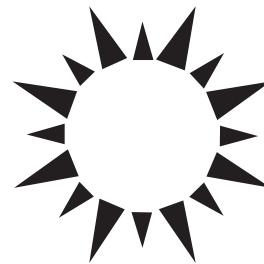
- Anonymous. (1965). Solar desalting gain acceptance in Greece. *Chem. Eng.* 72(26), 42–43.
- Belessiotis, V., Voropoulos, K., and Delyannis, E. (1996). Experimental and theoretical method for determination of daily output of a solar still: Input-output method. *Desalination* 100, 99–104.
- Bloemer, J. W., Collins, J. A., and Eibling, H. A. (1961, 1965, 1966). Solar distillation of saline water, Research and Development Progress Reports 50, 147, and 190. OSW, Washington, DC.
- Damm, W., and Geiser, P., *et al.* (1996). Wind powered MVC seawater desalination—Operating results and economic aspects. In “Proceedings of the Mediterranean Conference on Renewable Energy Sources for Water Production, June 10–17, Santorini,” pp. 143–146. Centre for Renewable Energy Sources, Athens, Greece.
- Delyannis, E., and Belessiotis, V. (1995a). Solar desalination: Is it effective? Part I. *Int. Desalination Water Reuse Q.* 4(4), 9–14.
- Delyannis, E., and Belessiotis, V. (1995b). Solar application in desalination: The Greek islands experiment. *Desalination* 100, 27–34.
- Howe, E. D., and Tleimat, B. D. (1967). Solar distiller to use in Coral islands. *Desalination* 2(1), 109–115.
- Delyannis, E., and Belessiotis, V. (1996). Solar desalination: Is it effective? Part II. *Int. Desalination Water Reuse Q.* 4(4), 28–34.
- Delyannis, E., and Belessiotis, V. (2001). Solar energy and desalination. *Adv. Solar Energy* 14, 287–330.
- Djebendjian, B., and Rayan, M. A. (2000). Theoretical investigation of the performance prediction of solar still: 4th International Water Technology Conference, Alexandria, VA, 1999. *Desalination* 128(2), 139–145.
- Duffie, J. A., and Beckman, W. A. (1991). “Solar Energy Thermal Processes.” Wiley, New York.
- Dunkle, R. V. (1961). Solar water distillation—The roof type still and a multi-effect diffusion still. *Int. Development in Heat Transfer, ASME Proc. Int. Heat Transfer, Part V, Univ. Colorado*, 895–902, 1961.
- Eckstrom, R. M. (1965). Design and construction of the Symi still. *Sun at Work* 10(1), 7.
- El-Nashar, A. M. (2000). Economics of small solar assisted multiple-effect stack distillation plants. *Desalination* 130, 200–215.
- El-Nashar, A. M., and Samad, M. (1996). A solar assisted seawater multi-effect distillation plant. 10 years of operating performance. In “Proceedings of the Mediterranean Conference on Renewable Energy Sources for Water Production, June 10–17, Santorini,” pp. 62–72. Centre for Renewable Energy Sources, Athens, Greece.
- Fath, H. E. S. (1996). High performance of a simple design, two effect solar distillation unit. *Desalination* 107, 223–233.
- Garcia-Rodriguez, L. (2002). Seawater desalination driven by renewable energies—A review. *Desalination* 143, 103–113.
- Glueckstern, P. (1999). Design and operation of medium and small-size desalination plants in remote areas. *Desalination* 122, 123–140.
- Goosen, M. F. A., Sablani, S. S., *et al.* (2000). Thermodynamic and economic considerations in solar desalination. *Desalination* 129, 63–89.
- Goswami, D. Y., Kreith, F., and Kreider, J. F. (2000). “Principles of Solar Engineering.” 2nd ed. Taylor & Francis, Philadelphia.
- Harding, J. (1883). Apparatus for solar distillation. *Proc. Inst. Civil Engineers London* 73, 284–288.
- Kreith, F., and Kreider, J. F. (1978). “Principles of Solar Engineering.” Hemisphere, New York.
- Morse, R. N., and Read, W. R. W. (1966). The development of a solar still for Australian conditions. In “Proceedings of the Conference on Power Production and Energy Conversion.” pp. 39–40. Institution of Engineers, Melbourne, Australia.
- Nebbia, G., and Nebbia-Menozzi, G. (1966). A short history of water desalination. In “Proceedings of the Aqua Dolce dal Mare, II Inchiesta Internazionale, April 1966, Milano.” pp. 129–172. Consiglio Nazionale delle Ricerche, Milano, Italy.
- Papanicolaou, E., Voropoulos, K., and Belessiotis V. (2002). Natural convection in a solar still. Accepted for publication.
- Robinson, N. (1955). Solar machines. In “Proceedings of the World Symposium on Applied Solar Energy, Phoenix, Arizona, November, 1–5,” pp. 43–61. Association of Applied Solar Energy, AZ, USA.

- Sardi, L. (1996). RO desalinators powered by PV systems for small/medium Italian islands. *In* "Proceedings of the Mediterranean Conference on Renewable Energy Sources for Water Production, June 10–17, Santorini," pp. 36–44. Centre for Renewable Energy Sources, Athens, Greece.
- Talbert, S. G., Eibling, J. A., and Loeff, G. O. G. (1970). "Solar Distillation of Saline Water Manual." Battele Memorial Institute, Columbus, OH.
- Telkes, M. (1945). Solar distiller for life rafts. U.S. Office of Technical Services Report 5225, MIT, OSRD Final Report to the National Defense Research Commission, 11.2, pp. 24–25. U.S. Office of Technical Services, Washington, DC.
- Telkes, M. (1955). Solar stills. *In* "Proceedings of the World Symposium on Applied Solar Energy, Phoenix, Arizona, November, 1–5," pp. 73–79. Association of Applied Solar Energy, AZ, USA.

Steel Production and Energy

MATTHIAS RUTH

University of Maryland
College Park, Maryland, United States



-
1. Background
 2. Iron and Steel Production
 3. Casting, Forming, and Finishing
 4. Ancillary Processes
 5. Drivers for Change of Energy Use in Iron and Steel Making
 6. Conclusions

Glossary

calcination A process of oxidation (here of lime in iron making).

charge Load of a furnace with the usual materials.

coke rate Amount of coke used per ton of iron or steel product.

energy efficiency Amount of output produced per unit of energy input.

energy intensity Amount of energy used per ton of product.

oxidation Chemical process by which oxygen is added to a reactant.

reduction Chemical process by which oxygen is removed from a reactant.

Iron and steel-based products lay the footprint for industrialization, and thus energy consumption in the transportation, commercial, industrial, and end-use sectors of an economy. Their manufacturing itself is highly energy intensive and often accounts for a significant share of a country's energy use. This article concentrates on the uses of energy for steel production.

1. BACKGROUND

Since the construction of the great Iron Bridge in England during the 1770s, the use of iron and steel-based products has come to be associated with the industrialization of economies. Iron and steel are used

in such diverse applications as railroads and skyscrapers, automobiles and other durable consumer products, as well as many infrastructure systems of modern society. To this day, iron and steel are among the least expensive and most widely used metals.

As economies industrialized and grew in size, world production of steel has increased (Fig. 1). By early 1900, the United States led the world in steel production, contributing about 36% of the world's output, followed by the German Empire (23%), the United Kingdom (17%), and Russia (10%). In 2000, the three largest steel producers were China (15% of world output), Japan (13%), and the United States (12%). With the boom in its economy, China's demand for finished steel increased more than 40-fold from the early 1950s to the mid-1990s, and its production rose by more than 10% per year. However, production increases in China were less rapid than the growth of the Japanese steel industry during that time period or the Korean steel industry since the mid-1970s.

This article explores the uses of energy for steel production. The first section describes key processes for raw steel production. The second section addresses iron and steel casting, finishing and forming processes used to produce semifinished or finished products. The third section describes processes that are not always carried out by the iron and steel industry itself, such as beneficiation of ores, but that significantly influence the industry's energy use profiles. The fourth section investigates various drivers behind technology change in the industry and their influences on energy use. The chapter closes with a brief summary and conclusions.

2. IRON AND STEEL PRODUCTION

Iron is made from ores by removing oxygen and impurities from the iron oxides in ores. Chemical reduction is achieved with the help of coke, lime, or

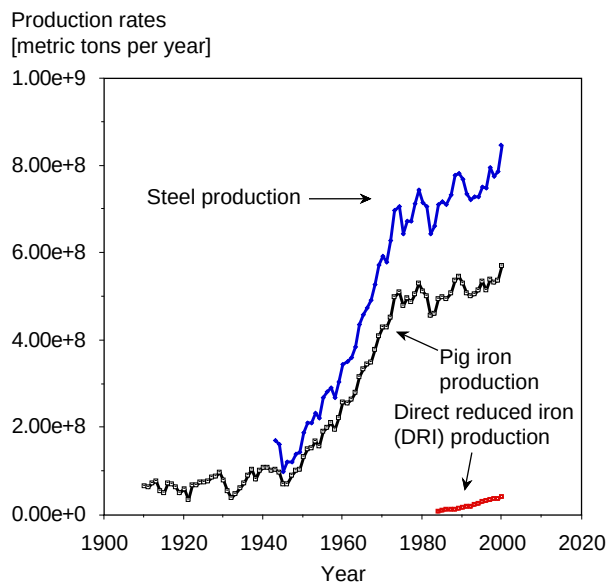


FIGURE 1 World production rates. Data from Kelly, T. D., and Fenton, M. D. (2002). Iron and steel statistics. U.S. Geological Survey, <http://minerals.usgs.gov/minerals/pubs/of01-006/ironand-steelslag.html>.

other reduction agents. The molten iron typically contains from 3 to 5% carbon and other impurities. It is usually produced with blast furnaces and passed on in a molten state as input in steel-making processes.

Steel making involves reducing carbon content to about 1%. Various steel-making processes exist. The crucible, bessemer, open hearth, and basic oxygen processes start with iron as their primary material input to which in some cases iron and steel scraps are added. In contrast, the electric arc process starts with up to 100% scrap as its material input and uses electricity to melt the furnace charge. Through history, each of these processes accounted for varying contributions to overall iron and steel production (Fig. 2). By early 1900 the crucible, bessemer, and open hearth furnaces were used in the United States for 67%, 32%, and 1% of steel production, respectively. By 2000, these furnaces were replaced in the U.S. steel-making process by the basic oxygen and electric arc furnaces, which produced 53% and 46% of steel, respectively.

Basic oxygen furnaces are typically part of vertically integrated operations that produce ores, a variety of intermediate products such as coke and pig iron, and raw and finished steels. While basic oxygen furnaces were replaced gradually in the United States and many other countries as the technology matured, furnaces aged, and new technologies became available, Japanese and European competitors had to

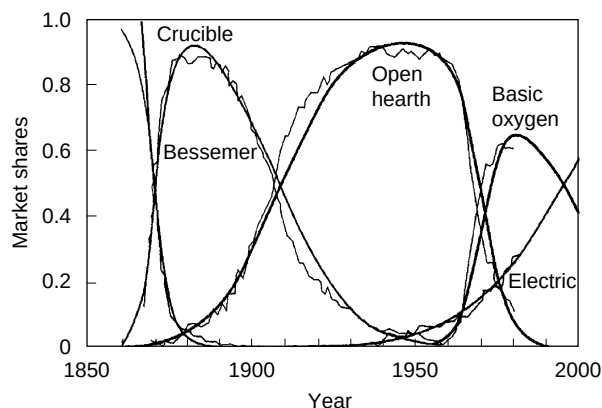


FIGURE 2 Process technology change in U.S. steel manufacturing, in fractional shares of raw steel tonnage produced. From Nakicenovic, N. (1990), Technological substitutions and long waves in the USA. In "Life Cycles and Long Waves, Lecture Notes in Economics and Mathematical Systems" (T. Vasko, R. U. Ayres, and L. Fontvieille, Eds.), No. 340, Springer-Verlag, Berlin. © Copyright Springer-Verlag. Permission granted.

rebuild most of their production facilities after the war. Only a small fraction of the furnaces they lost were replaced by the traditional open hearth technology. Most new capacity was added in the form of basic oxygen furnaces, which reduced the time it took to produce steel from longer than 6 hours to less than 40 minutes.

No new basic oxygen furnace capacity has been added since the 1960s in vertically integrated mills in the United States, capacity fell by over 50%, and no new future capacity additions are expected, in part because of large capital requirements to build furnaces that must operate at 1.5 to 3 million tons per year in order to be economically viable. Concomitant with these trends, the share of output from electric arc furnaces has steadily increased since the 1970s (Fig. 2), and the energy use profiles of the industry as a whole have changed.

Each iron and steel-making process has characteristic energy use profiles. However, in the thermodynamic ideal only iron making requires substantial energy consumption. The theoretical minimum energy requirement to reduce ore is given by the physical and chemical properties of ore. At room temperature, the theoretical minimum energy requirement to melt Hematite, the main ore type, is 6.9 GJ per ton of steel. This thermodynamic minimum does not change with time and establishes a lower bound on the energy necessary for steel production from ore. In practice, energy consumption for primary steel production is two to three

times the thermodynamic minimum. Consequently, energy use by the industry could, in principle, be reduced significantly.

2.1 Blast Furnace Production

Blast furnaces are the largest consumers of materials and energy in the iron and steel-making process. They are relatively flexible in their ability to use different metal charges, such as pellets, sinter, or scrap, with little change in performance. Other inputs include coke to produce carbon monoxide for the reduction of iron ore to iron, and other energy sources for various stages of the production process, such as preheating of air up to 1100°C for injection into blast furnaces. Since blast furnaces require significant quantities of coke, they are often operated in close conjunction with coke ovens, creating opportunities for joint use of byproducts, such as coke oven gas as well (Fig. 3).

The main outputs of blast furnaces includes pig iron and slag, which is formed by combination of

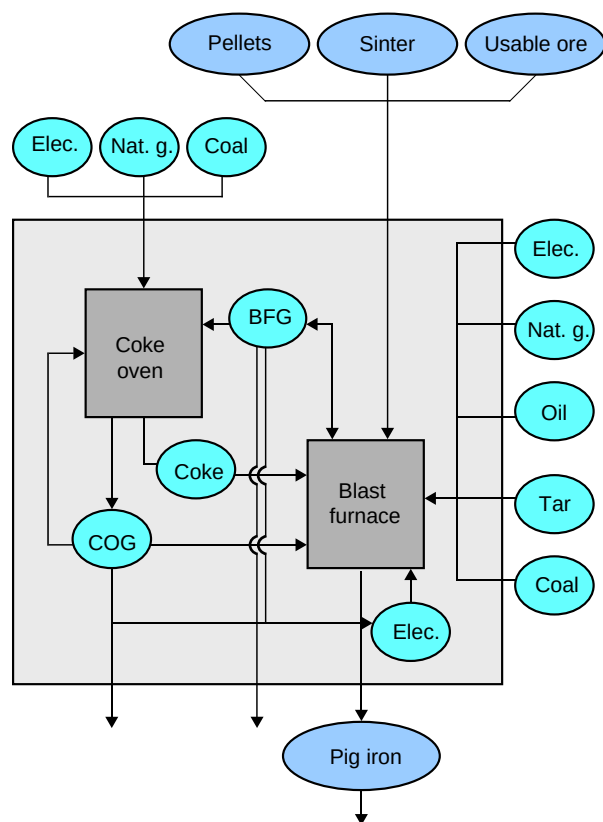


FIGURE 3 Main material and energy flows of blast furnace and coking operations. BFG = Blast furnace gas, COG, coke oven gas; Nat. G., natural gas; Elec., electricity.

limestone with sulfur and other impurities. Pig iron is typically tapped every 3 to 5 hours in quantities of 300 to 600 tons. After tapping, the pig iron is transported, typically in liquid form, to steel-making operations.

Other outputs from blast furnaces include blast furnace gas, an off-gas with a heating value of 3 MJ per m³, which is approximately one-tenth of the energy content of natural gas. This heating value has been dropping as a result of reduced coke rates and better burden preparation. The gas quality is insufficient to allow high-efficiency conversion to electricity in gas turbines. To increase heating values, blast furnace gas is often mixed with other gases such as coke oven gas and natural gas, which have higher energy content. This mixture can be converted into electricity with efficiencies of 41 to 42% in combined cycle plants. Much of that conversion takes place onsite at steel plants for electricity used in steel making and in some instances for sale into the electricity grid. A portion of the off-gas is consumed in boilers that produce compressed air via steam-powered or electric blowers. Funnelling that air through hot blast stoves provides the heat required in the blast.

With the price of coke rising, steel makers have sought to reduce their dependence on coke, for example, by installing pulverized coal injection into their blast furnaces as a partial substitute for coke. A few use natural gas for the same purpose. The need for greater efficiency in blast furnaces has also focused attention on the efficiency of oxygen use so that combining coal injection and large amounts of oxygen can further diminish the need for coke.

2.2 Crucible Steel Production

The crucible process for steel making was introduced in the 1740s in England. It used iron, frequently in solid form, which was molten with the aid of coke fires (Fig. 4). Melting took place in clay crucibles to which other materials were added to achieve desired properties of the steel they produced. Production was typically small-scale (approximately 50 kg per batch) and took roughly 4 to 6 hours per batch. Both the small batch sizes and relatively long heating times meant that the crucible process was highly energy inefficient and did not yield steel at rates necessary to meet rapidly growing demand.

2.3 Bessemer Steel Production

The bessemer process reduces molten pig iron in so-called bessemer converters—egg-shaped, silica, clay,

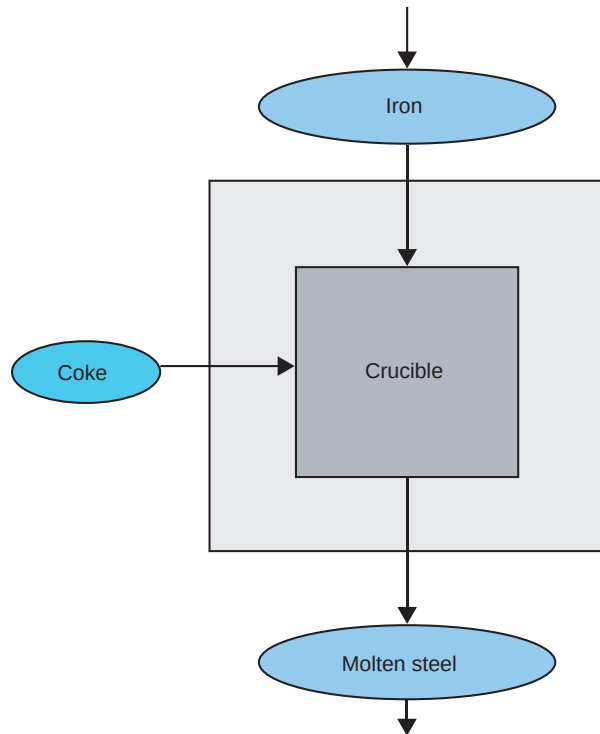


FIGURE 4 Main material and energy flows of crucible steel production.

or dolomite-lined containers with capacities of 5 to 30 tons of molten iron. An opening at the narrow upper portion of the bessemer converter allows iron to be introduced and the finished product to be poured out. Melting of metal typically is accomplished with coal and coke fires (Fig. 5). Air is forced upward through perforations at the wide bottom-end of the converter. As the air passes upward through the molten pig iron, impurities such as silicon, manganese, and carbon oxidize. Different property steels are produced by introducing additional elements, such as spiegeleisen—an iron-carbon-manganese alloy—into the molten metal after the oxidation is completed. The converter is then emptied into ladles from which the steel is poured into molds. Since the whole process could be completed typically in well under 1 hour and for significantly larger quantities than with crucibles, bessemer steel production significantly helped meet steel demands during the onset of the industrial revolution. As steel production increased, demand for coal and coke increased in lockstep.

2.4 Open Hearth Production

Open hearth furnaces are first charged with scrap and a small amount of limestone. Limestone is used

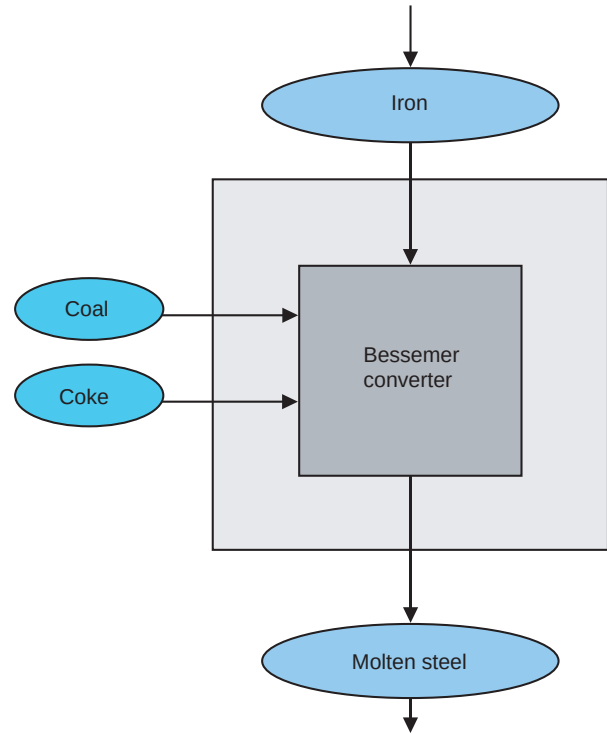


FIGURE 5 Main material and energy flows of bessemer steel production.

as a reduction agent. Then the fuel—typically oil, natural gas, coal, tar and pitch, or a combination of these—is ignited to melt the charge (Fig. 6). As the charge melts, proportioned amounts of molten pig iron are added to the furnace and high-purity oxygen is blown in. After various minor operations, the molten steel is trapped for approximately 8 hours to further chemically reduce the iron.

Between 1910 and 1970, open hearth furnaces were the predominant technology for steel production in much of the industrialized world. Their flexibility in using a wide range of hot metal and scrap charges added versatility to the production process. However, their long cycle times also meant significant energy use per ton of steel produced. With the onset of the energy crises of the 1970s and 1980s, the newer, more efficient technology of basic oxygen production became increasingly economically viable.

2.5 Basic Oxygen Furnace Production

Basic oxygen furnaces use pig iron and small amounts of scrap as their primary material inputs and typically melt the furnace charge with coke oven, blast furnace, and natural gas. Electricity and fuel oils are used to drive ancillary processes (Fig. 7).

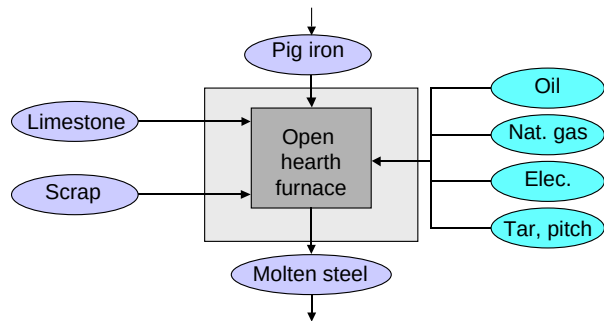


FIGURE 6 Main material and energy flows of open hearth steel production. Nat. G., natural gas; Elec., electricity.

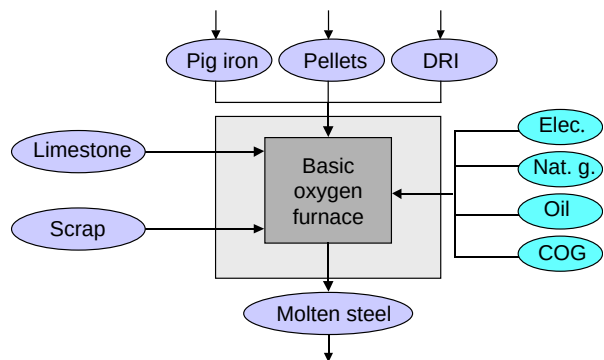


FIGURE 7 Main material and energy flows of basic oxygen production. DRI, direct reduced iron; COG, coke oven gas; Nat. G., natural gas; Elec., electricity.

Basic oxygen furnaces have the ability to cut total cycle time to less than a tenth of the cycle time of open hearth furnaces, thus enabling significant savings of energy for a given metal mix, and substantial total cost savings even though material costs themselves may be higher. However, their reduced flexibility in use of scraps, compared to the flexibility of basic oxygen furnaces, meant temporary scrap surpluses in the economy, reduced scrap prices, and opportunities for growth of yet another new technology—electric arc furnaces. At first, electric arc furnaces were predominantly used in integrated plants as the “swing” facility when scrap was scarce or plentiful. To a significant extent, this combination of versatile electric arc and efficient basic oxygen furnaces soon spelled the end for the open hearth as a major steel making technology.

2.6 Electric Arc Furnace Production

As output from integrated steel producers declined and electric arc technology improved, smaller-scale

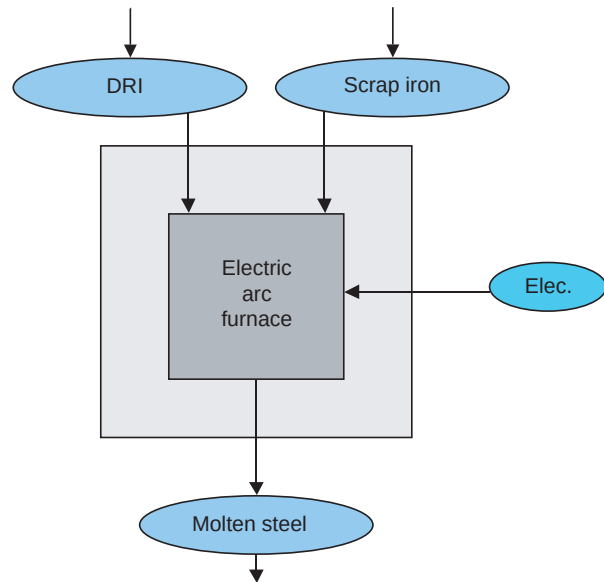


FIGURE 8 Main material and energy flows of electric arc furnace production. DRI, direct reduced iron; Elec., electricity.

electric arc furnaces quickly became an economically attractive alternative to integrated mills—compared, for example, to blast furnaces, which cost around \$1 billion and usually operate at scales between 1.5 and 3 million tons per year, as noted by Crandall in 1996.

Electric arc furnaces melt metal by passing a large electric current through the charge (Fig. 8). Since it takes approximately one-third to one-half of the energy to produce steel from scrap than from ore (because it takes considerably more energy to reduce iron oxides than simply melt scrap), energy savings and onsite emissions reductions are substantial when steel made from scrap is substituted for steel made from iron ore. Their relatively smaller scale, predominant use of scrap, and cleaner production enabled electric arc furnaces to be located in large urban areas and steel fabrication centers where the supply of steel scrap is high.

However, several limitations exist to reduce energy use in steel production by expanding electric arc furnace output. First, an increase of the amount of scrap that is recycled without some decrease in the amount of hot metal production will increase, not decrease, energy use. To reduce overall energy consumption for steel production requires substitution of either scrap for hot metal in existing furnaces, or of new, scrap-based furnaces for ore-based furnaces, not just simply an increase in scrap recycling. Second, actual energy savings depend on (1) how much energy is required to collect, sort, process, and deliver scrap and (2) how much energy

is consumed by hot metal production. Considering both the direct and indirect energy requirements, some studies have found that even if all the steel were made from scrap, the total energy intensity of U.S. steel production would be reduced by only 6%, saving less than 1% of total annual U.S. energy use. However, increases in scrap use may be pursued for other reasons, such as relative rises in labor costs and relatively high ore taxes. Other constraints on increasing the market share of electric arc furnace output depend on environmental cost, scrap availability and prices, and product quality.

3. CASTING, FORMING, AND FINISHING

A series of casting, forming, and finishing processes are used to change the shape and other physical and chemical characteristics of raw steel products. Each of these requires a specific set of steps with distinct efficiencies and energy use profiles. The general pattern of energy use of casting, forming, and finishing is shown in Fig. 9. Details are provided next.

3.1 Casting

Most of the raw steel produced by the industry is cast either continuously or into ingots or strips. In ingot casting, which in the United States and other industrialized countries is mostly used for small batch sizes for specialty steels or for end products with certain shape specifications, the molten steel is teemed into a series of molds and allowed to solidify to form ingots. After the molds are stripped away, the ingots are heated to uniform temperature in soaking pits to prepare them for rolling. Subsequently, heated ingots are removed from the pits and rolled into slabs, blooms, or billets.

In continuous casting, the molten steel is delivered in ladles and poured into a reservoir from which it is released into a mold. The steel cools as it passes through the mold and forms a solid outer skin. As the steel proceeds onto a runout table, the center of the steel also solidifies, yielding a semifinished shape at a specified width and thickness. Depending on the type of caster used, various shapes are produced. The melting, casting, and rolling processes have been linked with each other to produce shapes that substantially conform to the desired shape of the finished product.

Continuous casting eliminates the need for teeming into ingots, mold stripping, reheating, and

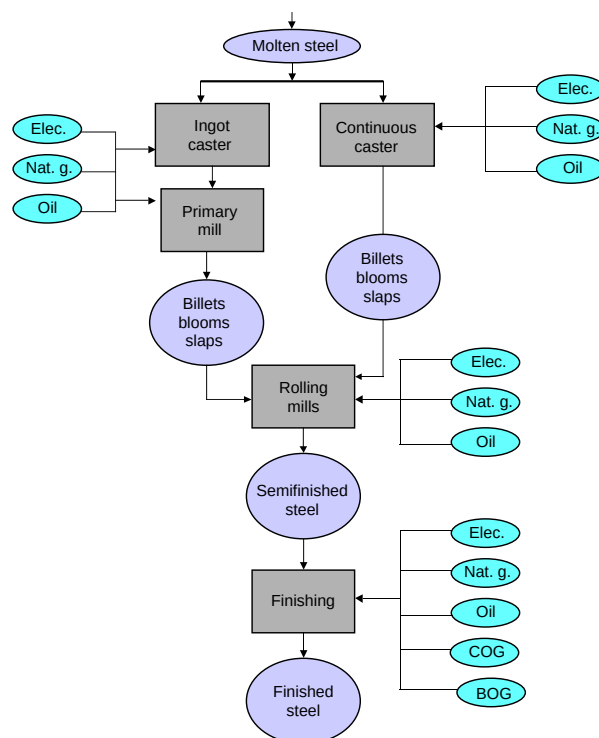


FIGURE 9 Main material and energy flows for pellet and sinter production. Nat. G., natural gas; Elec., electricity.

primary hot-rolling into semifinished shapes. As a consequence, the continuous process has higher yields compared to the ingot process and is roughly 10 times as energy efficient as continuous casting. Continuous casting also offers important benefits in product quality for most products because of the uniformity with which products are cast and the reduction of defects during solidification of the molten steel. Ultimately, the main benefit of continuous casting and direct rolling—both from an energy and overall economic perspective—may be the rapid feedback, which could eventually enable the elimination of most production errors that now result in substantial rejection of finished and semifinished products.

The choice of casting technology is closely related to existing melting technology. For example, in the United States in the 1970s, when open hearth and basic oxygen furnaces dominated steel production, less than 10% of steel was continuously cast. Similarly, in China and India where open hearth furnaces have still been widely used at the turn of the 21st century, the share of continuous casting was relatively low. In contrast, in the United States, approximately 96% of all steel produced at the turn

of the 21st century was continuously cast, the remaining 4% was ingot cast.

3.2 Forming

The slabs, billets, and blooms produced in casting operations are further processed to produce strips, sheets, plates, bars, rods, and other structural shapes through various hot forming operations, which are sometimes followed by cold forming operations. The exact sequence and temperatures of forming operations not only help generate desired shapes but also impart particular physical and chemical properties of the semifinished product. Prior to hot forming, semifinished shapes must be heated to rolling temperatures of about 1000 to 1200°C in a reheat furnace. Upon exiting the furnace, slabs may undergo a surface preparation step, which removes defects prior to entering a rolling mill.

The most common hot forming process is hot-rolling where a heated steel slab is passed between two rolls revolving in opposite direction. Each set of rolls produces an incremental reduction in the thickness of the slab. The final shape and characteristics of a hot formed piece depend on the rolling temperature, the roll profile, and the cooling process used after rolling. Rolling and reheating require considerable amounts of energy. The more the size of the cast steel resembles the size of the finished product, the less energy is required for finishing.

3.3 Finishing

Finishing processes are used to clean the surface of the semifinished hot-rolled steel products prior to cold-rolling or forming or coating operations. Rust and other oxides, oil and grease, and other impurities and contaminations of the steel are physically and chemically removed from the surface with help of solvent cleaners, pressurized water, abrasives, alkaline agents, salt baths, or acid pickling processes. The latter utilize steam generated from fuel to heat the pickling bath and preheat the incoming strip. Electricity is used to power various ancillary equipment, such as extraction fans and acid recovery. Acid recovery plants also use either steam or gas as primary energy inputs, depending on the type of acid employed.

Steel that has been hot-rolled and pickled may be cold-rolled to make a product thinner and smoother so that it is suitable for a variety of uses, such as car bodies, tin cans, pipes, and tubes. Energy requirements for rolling and finishing processes include fuels

used for slab reheating and heat treatment, and electricity used by rolling mills and cooling beds. The electricity requirement for rolling is determined by the initial and final gauges and quality of the steel. For hot-rolling it typically ranges from 600 to 1000 MJ per ton steel.

Reheat furnaces are fired with a variety of fuels, including natural gas, coke oven gas, basic oxygen furnace gas, and fuel oil. Modern, efficient reheating furnaces in hot strip mills use about 1400 MJ per ton of steel from cold start to heat slab, whereas older furnaces typically require around 1800 MJ per ton steel and most efficient technologies currently have energy intensities around 1000 MJ per ton.

In some flat-rolled electric arc furnace facilities, thin slabs are sheared and fed directly from the caster to an inline tunnel furnace that acts as a buffer ahead of the hot strip mill. These tunnel furnaces have energy intensities in the range of 500 to 800 MJ per ton steel.

The energy intensity of heat treating furnaces in plate mills is in the order of 400 MJ per ton. However, since not all plate steel is typically heat treated, the specific energy consumption of treated plates can be much higher.

Both electricity and fuel are consumed in cleaning and annealing. Modern hydrogen atmosphere annealing furnaces use 600 MJ per ton of product, while the energy intensity of older furnaces may be 50% or more higher.

4. ANCILLARY PROCESSES

4.1 Mining, Beneficiation, and Agglomeration

Since extraction and some preparation of ores are carried out at the mine site, not at the steel mill, they are typically considered outside the scope of steel making. However, considerable energy is expended in ore preparation processes (Fig. 10), and use of variously prepared ores can significantly reduce energy expenditures at the steel mill.

Crude ores are sold as usable ore by mine operations after some processing and beneficiation. One such form of usable ores are pellets. Pellets usually contain between 60 and 70% iron, with the remainder being binder and substances contained in the original ore. Their surface geometry enables more efficient heat and air fluxes through blast furnaces and thus more uniform reduction of iron oxides and higher efficiency in material (iron, coke) and energy

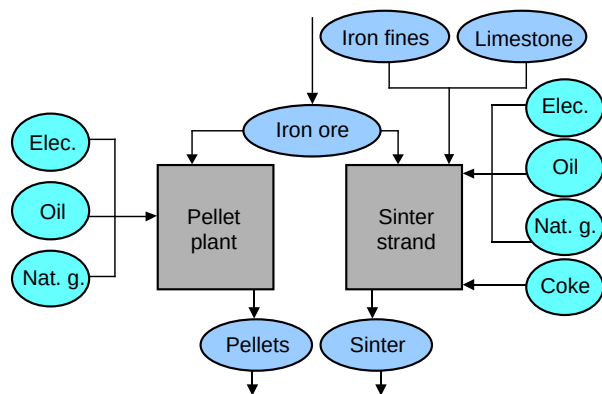


FIGURE 10 Casting, forming, and finishing processes. BOG, basic oxygen gas; COG, coke oven gas; Nat. G., natural gas; Elec., electricity.

use. Their production requires around 1000 MJ per ton of product, depending on the type of ore used, in addition to crushing, screening and washing of ores.

Other types of usable ores are distinguished, for example, as sinter and briquettes, depending on the process used. While pellet production typically occurs at the mine site or shipping port, sinter production usually occurs at the mill. A sintering plant provides a means of recycling many of the iron-bearing fines generated by iron making and steel making. In sinter strands, materials such as scale, sludge, and slag are added to ore fines and agglomerated into a suitable material. Flux, in the form of dolomite or limestone, is added to the sinter burden and calcined during the process, decreasing the energy requirements of the blast furnace process. The sinter, usually 50 to 60% iron, is easily reducible, since its high surface to volume ratio permits high contact of gases and solids. With energy intensities between 10,000 and 20,000 MJ per ton of product, sinter production is an order of magnitude more intensive than pellet production.

4.2 Coke Production

Coals suitable for coking are ground and blended before use as charge in so-called by-product coke ovens, which derive their name from the fact that during the coking process byproducts, such as light oils, tar, ammonia, and coke oven gas, are recovered for use elsewhere in the operation (see Fig. 3). Several of the byproduct fuels are mixed with combustion air for burning in combustion chambers. Under normal operating conditions, the charge is heated for 16 to 20 hours (14 or less with preheated coal—the latter

makes the process approximately 10% more efficient). Then the coke is forced into waiting cars where it is water quenched to prevent combustion. After cooling, the coke is crushed and screened. Most of it is then transported to blast furnaces and the remaining coke dust is conveyed to reclamation plants and used either in sintering plants or sold. The entire process requires upward of 20,000 MJ per net ton of coke.

Form coking, a complete substitute for the conventional slot oven coke process, involves grinding and preforming the coal, which is then binded, pelletized, cured to a char, and completely devolatilized to develop the necessary strength, a major problem in coke manufacture. Since briquettes are the final product, their uniform size and shape will improve blast furnace productivity. In addition, the process is enclosed, reducing emission problems. The process is not necessarily more energy efficient, although form coaking is a resource-saving technology.

Preheating of coal prior to use in the coking process may enable direct energy savings achieved in coke manufacture, improvements in the quality of the coke produced, and a consequent reduction in coke charged to the blast furnace. Alternatively, poorer quality coals may be used to produce coke equal in quality to that made from high-grade coals.

Because of high replacement costs, regionally declining supplies of low-sulfur coal and tightened emissions standards, coking capacity in the United States and many other industrialized countries has decreased significantly. Of the existing coke batteries, some are operated by iron and steel companies that produce coke mainly for their blast furnaces. The remainder is held by merchant companies that generally sell their product on the open market.

4.3 Scrap Preparation

Iron and steel scraps are used in many stages of iron and steel making. Some of the scraps (prompt scrap) accrue in production itself and are directly fed back into production, others (obsolete scrap) are returned to the industry from other industries or after use of iron and steel containing end use products. Typically, prompt scrap is of a more uniform quality and more readily available than obsolete scrap which often contains larger amounts of impurities and accrues after significant time lags between production and return to the industry.

Scrap quality is a limiting factor for increased scrap use. Certain types of scrap contain high

quantities of trace elements. High-quality steel requires a low trace element content. In the current practice, scrap is used for lower quality steel grades, such as reinforcement bars. However, electric arc furnace steel producers are gradually moving toward the production of higher quality products. In the longer run, declining steel scrap quality and increasing steel quality may pose serious challenges for recycling.

The lower the scrap quality, the lower the yield, and thus the higher the energy requirement per ton of desired output. Since prompt scrap is a waste product of the industry itself and does not replace raw material inputs, increases in prompt scrap use or quality will not noticeably improve energy efficiency. Only expansion of obsolete scrap use can lead to an overall reduction of energy requirements to the extent that obsolete scrap substitutes for steel production from ores. To increase productivity when using scrap and to replace high-cost electric energy with lower cost fuels, some basic oxygen furnaces and electric arc furnaces preheat scrap.

4.4 Direct Reduction

Direct reduced iron is iron ore in the form of lumps, fines or pellets that have had the oxygen removed by using hydrogen and carbon monoxide. Typical sources of carbon monoxide are natural gas, coal gas, and coal. Other energy inputs into the production process often include oil and electricity. Since much of the energy used is in the form of natural gas (e.g., Mexico produces virtually all its direct reduced iron with natural gas), most direct reduced iron is produced where low cost natural gas is available and where adequate capacities exist to access iron ore shipments. Use of natural gas not only reduces the need for coal and coke, but also reduces the associated emissions, such as sulfur oxides and carbon dioxide into the atmosphere.

In direct reduction the ore is heated to temperatures below its melting point. Direct energy requirements per net ton of direct reduced iron are in the order of 10,000 MJ. The product is 90 to 95% metallized and hence is suitable as a charge in blast, basic oxygen, and electric arc furnaces.

By using direct reduced iron, steel makers can dilute lower grade ores and lower cost ferrous scrap to attain acceptable levels of contamination in the mix. If direct reduced iron pellets are substituted for pelletized ores in blast furnaces, the amount of coke per ton of pig iron would be cut. However, reductions in the coke rate could result in shortages

of coke byproducts, while natural gas requirements would increase considerably.

5. DRIVERS FOR CHANGE OF ENERGY USE IN IRON AND STEEL MAKING

Changes in iron- and steel-making technology, capacity utilization and production quantities all influence energy use by the steel industry. These changes are often the result of a multitude of interrelated factors, including technological changes in steel production, changes in national and international markets of iron and steel and their substitutes, changes in availability and prices of inputs, as well as environmental regulation. As the industry meets the challenges associated with technology change, product market dynamics, changes in input prices, and environmental regulation, it tends to create savings in all factors of production: labor, capital, materials, and energy.

5.1 Technology Change

Among the most notable changes that occurred in the iron and steel industry of many industrialized countries since World War II are the transition from open hearth to basic oxygen and electric arc furnaces and the use of continuous casting methods instead of ingots. Many new steel-producing countries are following these trends as their industries expand, and technologies change. Further changes in the industry are expected as product quality from electric arc furnaces continues to improve and as newer technologies such as direct reduced iron provide increasingly cost-effective alternatives to the energy intensive basic oxygen route.

The extent to which iron and steel firms adopt new technology depends in part on previous successful experience with the technology, which makes innovation less of a gamble, and on the speed at which technologies change. Firms are likely to delay adoption of technologies that are improving rapidly and tend to adopt a technology after a slowdown in technical progress, either with the furnace the firm is choosing or the competing one. Since many improvements in technology are very costly, long-lived, and once implemented determine a firm's long-term material, energy, and emissions profiles, the industry's aggregate performance changes only slowly relative to the rate at which individual technologies change.

5.2 Market Dynamics

Prior to the 1970s, the industry experienced steady growth and sustained profitability. Since furnaces with higher capacity utilization have lower heat loss per ton of furnace charge and thus higher energy efficiencies and lower emissions per ton of product, overall energy efficiencies in the industry were relatively high during years of high production and capacity utilization rates.

The recession that followed the 1970s oil price crisis has led to overcapacity and resulted in pessimistic forecasts of future demand. Nevertheless, by the beginning of the 21st century, steel makers could produce 15% more steel than the world market demanded, but they were adding 2% more capacity each year.

Since the 1970s, many traditional steel firms have shed unprofitable operations, reorganized production to eliminate waste, and focused their resources both in terms of product lines and of the geographic markets they serve. However, U.S. integrated producers and in some cases minimills were slower than their European and Japanese competitors in implementing improvements.

Integrated steel producers were more likely to close plants that had comparably small output, competed with minimills or sold to contracting industries, and plants that did not use electric arc furnaces. As a consequence, the firms that remained in the market tended to be large, and the probability of exit fell as firm size increased. As the industry consolidated, far-reaching changes in technology required larger capital investments, and as market exit of firms decreased, rates of improvements in aggregate efficiencies from factors other than economies of scale slowed down.

Other reasons for delays in technology change included a lack of profits and access to capital at sufficiently low interest rates to make the necessary large investments in new steel-making technology. The situation was exacerbated by failures to anticipate changes in product needs in the rest of the economy and underestimates of the potentials of new technologies to reduce cost and increase product quality.

5.3 Changes in Material and Energy Prices

Changes in quality of material inputs into steel making are closely related to changes in energy requirements to process furnace charge. As a

consequence, the material and energy costs of steel production are related to each other as well as to other technological variables and firm experience.

For example, availability and quality of iron ores influence energy and material efficiencies. Iron ores used in China are mainly of low grade with an average iron content of 34%. Iron ores are beneficiated before they are put into blast furnaces, but because of limits to which the quality of ores can be improved through beneficiation, the Chinese steel industry historically had high coke rates and high iron-to-steel ratios. In contrast, in the 1970s and 1980s, Japanese firms have secured high-quality ore at favorable prices by engaging in long-term contracts with Brazilian and Australian ore development projects. During that time, the price of ore had dropped on the international markets to the point where Japanese firms could secure ores at prices well below the cost incurred, for example, by American firms to get ore from domestic sources, even though shipping cost for Japanese firms were substantial. The resulting cost savings helped Japanese firms to increase their market shares and profitability and thus afforded capital for investment in new capacity and improved technologies.

Improvements in competitiveness and technology also occur in response to changes in energy prices, although influences of energy prices—and for that matter any other single driver—may have been lower than often suspected. Empirical results show that even though energy is a large cost component, energy prices are not a central criterion by which integrated or minimills decide about major process changes. Other determinants include prices and quality of reduction agents. For example, for the U.S. iron and steel industry as a whole, higher coke prices are correlated with later adoption of basic oxygen furnaces and earlier adoption of electric arc furnaces.

Further influences on technology choice and energy efficiency may come from investments into a more diversified and reliable energy system, demand side management, and technology-specific investments. For example, improvements in natural gas infrastructure and technology, power recovery, and direct coal injection at the blast furnace have increased energy efficiency in the steel industry of China and its economy as a whole. Potential improvements in minimills' electricity use indicate that minimills may be a particularly good target for energy demand policies. They are large single users of electricity with a demonstrated engineering potential for efficiency improvements. Encouraging

more rapid upgrading or replacement of older capacity is one effective way to improve aggregate energy efficiency.

5.4 Environmental Regulation

Many energy intensive industries like steel are under increasing pressure to reduce their energy consumption for at least the following two key reasons. First, improving the efficiency with which energy along with other inputs is used can help realize operating cost savings and thus contribute to the industry's competitiveness. Second, reductions in energy use may occur for environmental reasons because environmental impact may reduce the appeal of steel relative to competing materials or because the need to comply with environmental regulation may increase the cost of production.

Focus on local and regional water and air quality has led to the closure of many coke ovens and smelters whose needs for cooling water and emissions of sulfur oxides resulted in significant environmental impacts. In many industrialized countries, blast furnaces continue to face closure and steel makers are increasingly reluctant to make long-term maintenance expenditures on blast furnaces and their associated technologies.

Environmental concerns have shifted to emissions of carbon dioxide and other greenhouse gases influencing global climate. With its use of large quantities of fossil fuels, the iron and steel industry accounts for about 7% of total anthropogenic carbon dioxide emissions. Efforts to reduce carbon dioxide emissions through fuel switching and efficiency improvements typically also address other inefficiencies in the industry.

A number of fuel alternatives exist for coal, the predominant source of carbon dioxide emissions by the industry. The most relevant ones from a carbon dioxide reduction perspective are natural gas, charcoal from renewable wood biomass, waste plastic, hydrogen, and electricity.

Natural gas can be injected into blast furnaces and with proper oxygen enrichment can boost hot-metal output by up to 40%, compared to about 10% in the case of coal injection. Up to 125 kg of natural gas can be injected per ton of hot metal (in energy terms equivalent to about 210 kg coal injection). The potential for carbon dioxide emission reductions amounts to 100 Mt (compared to coal injection) or about 135 Mt compared to coke use. Low carbon dioxide emissions from Mexican steel producers have in part been achieved by using a larger

proportion of natural gas in the steel-making process compared to more carbon-rich fuels.

Charcoal, if produced from wood can be considered a renewable carbon source. Part of Brazilian iron production is based on the use of charcoal. Charcoal lacks the mechanical stability of coke, but it is possible to substitute blast furnace coal injection with charcoal injection. Traditional charcoal production is a very polluting activity. New, modern shaft kilns have been developed with much lower emissions compared to traditional charcoal making, but production costs are significantly higher.

Plastic wastes can be added to coke ovens to reduce coal requirements or injected into blast furnaces to substitute for coke and coal. An advantage of use in coke ovens over injection into blast furnaces lies in the possibility to generate higher value products and reduced interference with the sensitive blast furnace operation. However, byproducts from incineration of plastics, most notably gaseous emissions and generation of toxic ash, may limit overall use of plastic wastes by the steel industry.

Not all changes in fuel mix reduce overall carbon dioxide emissions. As firms in different countries undertake fuel substitution in efforts to reduce operating costs, some increases in carbon dioxide emissions may result. For example, the fuel mix in Brazil's steel industry is less carbon dioxide-intensive than in other major steel producing countries because the share of coal is low. The observed increases in net carbon dioxide emissions from Brazilian steel making have been triggered, in part, by the privatization of the steel industry in Brazil which led to a trend away from wood-based charcoal and capital-intensive hydropower generation to fossil-based electricity.

6. CONCLUSIONS

Iron and steel production is among the most energy-intensive industrial activities. Significant changes occurred in the dominant technologies used to produce steel, as well as the fuel mix of the industry and the efficiency of each technology. Several technologies widely used at the beginning of the past century—open hearth and basic oxygen furnaces—now account for only small market shares.

As technology mix in the iron and steel industry changed, upstream and downstream processes had to

adjust. Pelletization of ores and preheating of scraps, for example, are increasingly widely used methods to improve furnace performance, material and energy efficiency, and product quality. New casting and rolling technologies similarly help improve overall industry performance. For example, thin-slab casting in minimills eliminates the need for traditional hot-rolling facilities to reduce slab to hot-rolled sheet. Similarly, newly developing thin-strip casting may produce light-gauge sheet steel without several hot and cold reduction steps.

New technologies will continue to enable minimills to enter new markets. Minimills will also increase the usage of some form of iron—direct reduced iron, pig iron, hot metal—with the scrap charge to improve the quality of the steel they produce to serve even broader markets.

As the industry responds to changes in energy and material prices, changes in the cost of capital, market dynamics, and environmental regulation, the ratio of energy consumption by the steel industry to gross national product declines. For example, in constant-dollar terms, that decline was 32% in the United States between 1973 and 1984. In western Europe and Japan, input prices have risen even more sharply and even more energy conservation has been carried out.

Further improvements in technologies undoubtedly will continue to help maintain steel's competitiveness and energy and environmental performances. Many of these improvements will support the transition from aging furnaces in integrated mills to modern, smaller, more efficient operations. These changes may be most striking in China, India, and other developing countries with significant steel-production capacities and rising demand for iron- and steel-based products.

SEE ALSO THE FOLLOWING ARTICLES

Aluminum Production and Energy • Cement and Energy • Early Industrial World, Energy Flow in • Glass and Energy • Material Efficiency and Energy Use • Material Use in Automobiles • Plastics Production and Energy • Recycling of Metals • Reuse and Energy

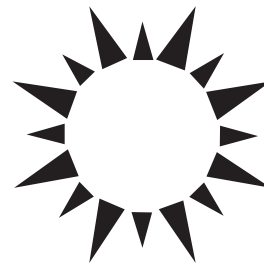
Further Reading

- Ahlbrandt, R. S., Fruehan, R. J., *et al.* (1996). "The Renaissance of American Steel." Oxford University Press, New York, Oxford.
- Crandall, R. W. (1996). From competitiveness to competition: The threat of minimills to large national steel companies. *Resources Policy* 22(1/2), 107–118.
- Fenton, M. D. (2000). Iron and steel, U. S. Geological Survey.
- Gielen, G., and Moriguchi, Y. (2002). CO₂ in the iron and steel industry: An analysis of Japanese emission reduction potentials. *Energy Policy* 30, 849–863.
- Hannon, B., and Brodrick, J. R. (1982). Steel recycling and energy conservation. *Science* 216, 485–491.
- Kim, Y., and Worrell, E. (2002). International comparison of CO₂ emission trends in the iron and steel industry. *Energy Policy* 30, 827–838.
- Office of Industrial Technology (OIT) (2000). Steel: Industry of the future. Washington, D.C., Office of Industrial Technology, U.S. Department of Energy, 98.
- Ruth, M. (1995). Technology change in U.S. iron and steel production: Implications for material and energy use, and CO₂ emissions. *Resources Policy* 21(3), 199–214.
- Ruth, M., Amato, A., and Davidsdottir, B. (2000). Impacts of market-based climate change policy on the U.S. iron and steel industry. *Energy Sources* 22(3), 269–280.
- Sparrow, F. T. (1983). "Energy and material flows in the iron and steel industry," p. 107. Argonne National Laboratory, Argonne, IL.
- Sugimoto, T. (1993). The Chinese steel industry. *Resources Policy* (December), 264–286.
- U.S. Geological Survey (USGS) (2002). Mineral commodity summaries, 86–87.

Stock Markets and Energy Prices

PERRY SADORSKY

York University
Toronto, Ontario, Canada



-
1. Definition and Economic Importance
 2. Relationship between Stock Markets, Energy Prices, and the Economy
 3. Oil Markets and Oil Prices
 4. Futures Markets, Oil Price Volatility, and Stock Markets
 5. Oil Prices and Currency Exchange Rate Risk
 6. Technology Stock Prices and Energy Prices in the New Economy
 7. Modeling Approach and Quantitative Methods
 8. Empirical Evidence on the Relationship between Oil Price Movements and Stock Price Movements
 9. Energy Price Risk Management

Glossary

- arbitrage pricing theory (APT)** A theoretical development in corporate finance that yields a model for pricing risk. The APT model linearly relates stock returns to several risk factors.
- autoregressive conditional heteroskedasticity (ARCH)** A statistical modeling technique that takes into account volatility clustering.
- backwardation** A pricing structure in which futures prices are lower than spot prices or when the near futures price exceeds the distant futures price.
- contango** A pricing structure in which futures prices are higher than spot prices.
- forward contract** A contract for the purchase or sale of a commodity for future delivery, given by a specific delivery date, and not traded on an organized market.
- futures contract** Standardized contract for the purchase or sale of a commodity for future delivery, given by a specific delivery date, and traded on an organized futures markets.
- heteroskedasticity** The variances of a set of random variables are not constant.
- ordinary least squares (OLS)** A statistical modeling technique that posits a linear relationship between a set of drivers and a dependent variable.
- risk** A source of randomness that can be quantified.

uncertainty A source of randomness that cannot be quantified.

vector autoregression (VAR) A statistical modeling technique in which each variable of interest is explained by lagged values of itself and the lagged values of other variables in the system of equations.

volatility A tendency for prices to fluctuate widely.

volatility clustering A tendency for large movements in prices to be followed by large movements in prices in either direction.

The interaction between changes in energy prices and economic performance has been of interest for a long time. For example, in the 1890s J. D. Rockefeller publicly expressed concern over rapidly fluctuating oil prices and advocated controlling oil prices in order to stabilize oil company profits and economic activity. Conversely, the relationship between stock markets and energy prices is a relatively new and emerging field of energy economics and finance that sprung to life after the two oil price shocks of the 1970s and the development of the oil futures market in the 1980s. Serious study of the relationship between stock markets and energy prices, and energy and the macroeconomy, has only occurred during the past 20 years.

1. DEFINITION AND ECONOMIC IMPORTANCE

Higher energy prices, mostly through higher oil prices, act like an inflation tax on consumers and producers. Higher oil prices raise the cost of production for non-oil-producing companies and, in the absence of fully passing these costs on to consumers, reduce profits and dividends. Profits and dividends are included in the income side of a country's national accounts. Consequently, sharp reductions in profits and dividends reduce gross

domestic product (GDP). Reduced profits dampen the future outlook for corporate earnings, a major driver of stock prices, and decrease stock returns. Steel mills, paper mills, sawmills, airlines, trucking firms, and other companies lay off workers and close plants when faced with steadily increasing fuel bills. Higher oil prices impact consumers directly via increases in the price of gasoline. This reduces the amount of disposable income consumers have left for purchasing other goods and services. Firms selling nonessential goods and services experience a decrease in business activity, which in turn results in lower profits, corporate earnings, and dividends and feeds back into the economy and the stock market.

With the U.S. Federal Reserve moving to an ever more transparent monetary policy, oil price changes have the potential to have a larger impact on stock prices than interest rate changes. This is because oil prices fluctuate more than interest rates do, and oil price changes embody more uncertainty regarding oil demand and supply imbalance, geopolitical, and institutional factors than do interest rates.

2. RELATIONSHIP BETWEEN STOCK MARKETS, ENERGY PRICES, AND THE ECONOMY

Real crude oil prices were highest in two distinct periods: 1861–1869 and 1974–1981. Real oil prices in the period 1861–1869 were very volatile and this is historically identified as the period with the highest real oil prices and the most amount of oil price volatility. Russian oil exports began in the 1880s and this helped to lower oil prices. After the discovery of oil at Spindletop, Texas, at the turn of the 20th century, real oil prices tended to stay fairly constant for long periods of time, particularly between the 1930s (at approximately the time that the East Texas oil field was discovered) and the early 1970s. Price changes did occur, but these changes were few and far between. After the oil price shocks of 1973 and 1979, the underlying structure of the world oil market changed and oil price volatility increased considerably. In fact, oil price volatility in the late 1970s was almost as high as that during the 1860s. Increased oil price volatility created a need for risk management tools and the development of forward and futures markets for crude oil followed. Futures markets for natural gas and electricity are also important, but most of the attention regarding stock markets and energy prices is focused on the crude oil

futures market. Today, it is the interaction between oil futures markets and the stock market that financial managers watch closely for signs of pricing pressure on the economy. In the future, with deregulated natural gas and electricity markets, the prices of natural gas and electricity will have a much larger impact on financial markets and the economy.

Since World War II, almost every recession in the United States has been preceded by a large increase in oil prices. Generally, the negative correlation between oil prices and GDP is taken as an empirical fact. During the four quarters preceding the onset of the 1948–1949, 1953–1954, 1957–1958, 1960–1961, and 1969–1970 contractions, the real price of energy increased on average 1.5%. The average increase in real energy prices prior to the onset of the four recessions during the periods 1973–1975 (called the U-shaped recession because real GDP growth during this period resembled the letter U), 1980, 1981–1982 (the W-shaped recession), and 1990–1991 (the V-shaped recession) was 17.5%. This increase in real oil prices was much greater than in earlier recessions. The economic downturns in the United States during the 1970s and 1980s were the largest since the Great Depression. The impact of oil price shocks on the U.S. economy became more pronounced after 1970. This was the time when U.S. demand for imported oil increased and the Organization of Petroleum Exporting Countries (OPEC) cartel began to exert their influence on the world oil market. The 1970s and early 1980 were also turbulent times for world equity markets. Oil price shocks can also occur from large decreases in oil prices. Large decreases in oil prices in 1986 and 1998 are two such examples.

Until the early 1970s, the price of internationally traded oil was set by a group of major oil companies (known as the seven sisters). During this time, the major oil companies set the price of oil on a reference barrel of oil (Arabian Light 34 API) so that the price remained constant in nominal dollars. OPEC basically inherited this pricing regime with some modifications. It was not until 1987 that OPEC adopted the current pricing regime whereby the price of internationally traded oil is determined by a pricing formula based on a reference basket of crude oils.

To understand the relationship between stock markets and energy prices, it is useful to use an equity pricing model. In such a model, the price of equity at any point in time is equal to the expected present value of discounted future cash flows. A firm's cash flow can be affected by economy-wide

factors as well as firm-specific factors. Approximately 60–80% of stock price movements result from broad-based market and industry sector factors and only 20–40% are related to company-specific factors. Cash flows are affected by business cycle conditions, and business cycle conditions can be represented by several key macroeconomic variables, such as industrial production, crude oil prices, interest rates, consumer prices, and possibly currency exchange rates. The ability to manage firm-specific and economy-wide risks determines the economic evolution of a firm.

Oil prices can affect the expected present value of cash flows in several ways. Oil is a resource and essential input into the production process. The basic production function requires inputs on capital (K), labor (L), energy (E), and materials (M): This is the basis for the KLEM model. Increases in energy costs, with the prices of other inputs held constant, will increase the cost of production and, without any offsetting effects, will reduce company profits and dividends. More generally, energy price increases can be mitigated by substitution effects among the other factor inputs. The impact on stock prices depends on whether the company is a net producer or consumer of energy. Increases in energy prices lower (raise) the stock prices of energy-consuming (producing) companies.

The net effect of higher energy prices on the stock market depends on how many firms are net users of oil. Given that there are more firms using oil than producing oil, oil is an input to the global economy. As a result, increases in oil prices lead to decreases in stock prices.

Future oil price changes can also impact stock prices through the discount rate. The expected discount rate is composed of a real interest rate component and an expected inflation component. Higher oil prices put upward pressure on the general price level, thereby raising the aggregate supply curve for a country that is a net user of oil and oil-related products. Higher oil prices can lead to increases in expected inflation and higher nominal interest rates. Since interest rates are negatively correlated with stock prices, increases in interest rates depress stock prices.

Economic growth leads to feedback effects between the economy, energy prices, and the stock market. Increasing industrial production generates a stronger economy and, as a consequence, higher profits and dividends. Higher profits and dividends raise stock prices. A stronger economy also raises the demand for oil and oil-related products. An increased demand for oil and oil-related products puts

upward pressure on aggregate prices and interest rates, and the reaction of the Central Bank becomes critical to the movement of the stock market. The Central Bank can either accommodate or dampen the increased economic activity via expansionary or contractionary monetary policy, respectively. Which path the Central Bank takes directly impacts stock returns.

Interest rates can impact stock prices in three ways. First, changes in interest rates are changes in the price charged for credit. The price of credit affects both firms and consumers. Firms seeking to expand existing operations or build new ones need to borrow money. Consequently, the price charged for credit is an important influence on the level of corporate profits and changes in corporate profits are a major driver of stock prices. The price of credit, through its impact on corporate profits, affects the price that investors are willing to pay for equities. Companies prefer to borrow money or issue debt when interest rates are low. In addition, a higher interest rate makes it less likely for consumers to purchase big ticket items, such as houses, automobiles, furniture, and appliances. Second, higher interest rates alter the relationship between competing financial assets. As interest rates rise, fixed-income securities become more attractive relative to equities and investors rebalance their portfolios to take this into account. Third, some stocks are purchased on margin and changes in interest rates impact the cost of carrying margin debt. Higher margin debt limits the ability of investors to participate in the stock market.

Inflation, a sustained rise in the general price level, erodes the real value of investments. This was particularly evident in the high-inflation period of the 1970s and early 1980s. As an example, consider the case of person A loaning \$100 to person B for 1 year. Person A would like to earn a 3% real rate of return. Suppose that the current inflation rate is 2% per year and this is expected to be the inflation rate in the following year. Person A will then loan person B \$100 at an interest rate of 5%. At the end of 1 year, person B agrees to pay person A \$105. Now suppose that the inflation rate over the coming year is not 2% but instead 4%. In this case, person A's real return on his or her investment is 1% ($5\% - 4\%$). Inflation (unexpected) has lowered the return on person A's investment.

Furthermore, inflationary pressures tend to lead to higher interest rates and a slowing economy, which lessens the ability of growth companies to meet revenue and earnings targets, causing stock prices to

fall. Oil price increases are one of the most likely sources of inflation and oil is a raw material used in almost everything that is made. Crude oil and natural gas are required to produce 98% of North American manufactured products. Oil price increases usually impact the transportation sector first through higher prices for fuel (gasoline, diesel, and jet fuel), with initial price hikes being absorbed by this sector. Airlines are particularly susceptible to higher oil prices because jet fuel represents a major cost of doing business in the airline industry. Persistent oil price increases lead to higher producer prices and eventually, after 6–9 months, inflation. Inflation, which erodes the real value of investments, leads to higher interest rates as contractionary monetary policy is used to bring economy-wide price increases down. Crude oil price increases in 1973, 1979, and 1990 were followed by stock market downturns.

Although oil price increases are not very good for the performance of the economy as a whole, oil price increases are very beneficial to companies engaged in the oil and gas industry. Consequently, oil price risk is a major concern for oil and gas companies. Higher oil prices lead to increased profits and spending on capital budgets, whereas lower oil prices lead to lower cash flows and reduced spending patterns of oil producers. Annualized oil price volatility is approximately 25% per year and annualized natural gas price volatility is approximately 40%. Hence, price variability is a major impact on the stock returns of oil and gas companies. There are three ways in which crude oil prices affect oil company stock prices. First, investors in oil stocks are usually only concerned about the flow of profits from their stock investment and future profits are determined by the expected average price of oil over a particular period. If investors expect the average price of oil to fall (rise), then they will sell (buy) shares in oil companies. Second, oil company stocks are more leveraged than crude oil. Oil exploration, development, and production is a capital-intensive business and oil and gas companies often borrow large amounts of money to go into production. The capital costs of exploration, development, and production depend on several factors, including whether the deposit is underground, under the ocean, or open pit; the type of plant required for oil extraction; and the location of the plant. As oil prices rise or fall, marginal oil deposits become profitable or unprofitable, which in turn changes the value of the company by more than the change in the actual price of crude oil. Third, the profits of oil companies, as for other companies, are a function of production

costs and production costs are influenced by the economic business cycle, which in turn is affected by energy prices. In the usual case, rising energy prices contribute more to the revenue side of an energy company's profits than to the cost side.

3. OIL MARKETS AND OIL PRICES

The global oil and gas industry is a mature industry that has prospects for growth as long as global demand for oil and gas products continues to grow. Oil and gas product demand is difficult to predict but is generally highly correlated with worldwide industrial production. In 2000, world oil consumption was 73,905 thousand barrels of oil per day (b/d). The United States accounted for 25.6% of this world total, whereas Europe accounted for 21.4%. During the period 1990–2000, world oil consumption increased 12.9%. World demand for oil is expected to grow 2% per year during the next decade. United States energy demand sensitivities indicate that a 1% increase in real U.S. GDP raises U.S. petroleum demand by approximately 0.6% and U.S. natural gas demand by approximately 1.1%. A 10% increase in crude oil prices, assuming no price response from nonpetroleum energy sources, reduces U.S. petroleum demand by 0.3%. These data were obtained from the U.S. Energy Information Administration Short-Term Energy Outlook (www.eia.doe.gov/emeu/steo/pub/sensitivities.html). The oil and gas industry, however, is very sensitive to business cycle conditions.

In 2000, world oil production was 74,510 thousand b/d. The five largest producing countries were Saudi Arabia (12.3% of world production), the United States (9.8%), the Russian Federation (9.0%), Iran (5.2%), and Mexico (4.8%). In 2000, OPEC accounted for 41% of world production. When OPEC member countries act in unison, they command monopoly power over approximately one-third of the world's oil production. During the period 1990–2000, world oil production increased by 13.9%.

Of course, future oil production depends on where the oil reserves are located. This is where it becomes interesting because most of the world's consumption of oil (62.4% of the world total) occurs in the Organization of Economic Cooperation and Development (OECD) countries, whereas most of the world's oil reserves are located outside of OECD countries. In fact, OECD countries only account for 8.1% of the world's oil reserves. For example, the United States, the world's largest consumer of oil,

currently imports approximately 60% of its crude oil. During 2001, approximately 48% of U.S. crude oil imports came from the Western Hemisphere (19% from South America, 15% from Mexico, and 14% from Canada), whereas 30% came from the Persian Gulf region (18% from Saudi Arabia, 9% from Iraq, and 3% from Kuwait). Since 1960, U.S. petroleum imports have increased more than sixfold.

In 2000, total world proven oil reserves were estimated at 1046.4 thousand million barrels of oil. At current production rates, this implies that there are 39.9 years of oil remaining in the world. Saudi Arabia has the most proven oil reserves with 261.7 thousand million barrels of oil. This equals 25% of the world total. OPEC countries own 77.8% of the world's proven oil reserves. In comparison, the United States and Mexico have proven oil reserves numbering 29.7 (2.8% of the world total) and 28.3 (2.7%) thousand million barrels of oil, respectively. Plentiful cheap oil is found mostly in OPEC countries. The wild card is the Russian Federation, which, in the interim, can rival Saudi Arabia in terms of crude oil production.

In a competitive well-functioning market, the price of oil would be determined by the intersection of demand and supply. An increase in demand (assuming supply does not change) would increase oil prices and an increase in supply (assuming demand does not change) would decrease oil prices. The actual mechanics of the world oil market, however, are much more complicated and are heavily influenced by global politics, institutions, and futures markets.

Political factors are important determinants in oil policy in both the oil-exporting countries (historically OPEC) and the oil-importing countries. The oil-exporting countries tend to intervene in the oil market when prices are too low. Low oil prices reduce the revenue of oil-exporting countries, dampening their economic growth and prosperity. OPEC sells oil in international markets on the basis of price formulae that use as a reference the spot or futures prices of certain important crude oil prices, such as West Texas Intermediate, Brent, or Dubai. Oil-importing countries often enact oil-directed energy policies when oil prices become too high. Higher oil prices tend to be followed by inflation, which erodes the real value of investments and slows economic growth and well-being. Oil-exporting countries can also use oil as a strategic weapon in response to U.S. foreign policy. The world oil market is populated by many consumers and many suppliers, which by itself suggests a competitive market for crude oil. The

reality is that the world oil market is very much influenced by institutional factors such as OPEC, strategic oil reserves in OECD countries, and the role that the large multinational oil companies play.

Generally, the difference between oil that is produced and oil that is consumed, if positive, goes into oil inventory stocks. Short-term supply disruptions can be met with a drawdown of oil stocks. Obtaining reliable estimates on oil stocks, however, is difficult because different countries use different reporting systems. Moreover, there is the issue of oil tanker traffic, which plays a major role in getting oil out of the producing countries in the Middle East and into the OECD consuming nations. Supply bottlenecks caused by oil tanker traffic can have a huge impact on gasoline prices in oil-importing countries.

Trade on the marginal barrel of oil sets the price of crude oil. The 11 members of OPEC (Algeria, Indonesia, Iran, Iraq, Kuwait, Libya, Nigeria, Qatar, Saudi Arabia, United Arab Emirates, and Venezuela) are the marginal producers. The near-term outlook for the oil market can be determined by matching data on future consumption, OPEC oil being transported by tanker, and oil inventories. The problem is that OPEC members are very secretive about how much oil they are actually exporting at any given time—so secretive, in fact, that they do not even reveal their oil exports to each other. Tracking the amount of oil in transit is difficult but essential to understanding the world oil market.

The bulk of OPEC oil moves from Quoin Island in the Persian Gulf to ports in New York, Gibraltar (via the Suez Canal), New Orleans, Montevideo, Uruguay, Fremantle, Australia, and Yokohama, Japan. It can take several days to fully load an oil tanker at Quoin Island. Once an oil tanker is loaded, it sails to its destination. Sailing time depends on the distance traveled. The shortest trips of 12 or 13 sailing days are from Quoin Island to Gibraltar and Fremantle. The longest trip takes 31 days to New Orleans. Once the crude oil arrives at its destination, it is unloaded and piped to refineries to be turned into refined products such as gasoline, heating oil, and jet fuel. The refined products are then moved in pipelines and trucks to consumer outlets. This process takes approximately 2 months. In total, it can take 3 or 4 months for crude oil from the Persian Gulf to reach gasoline stations in the United States. Consequently, knowing the amount of oil in transit is essential to knowing the amount of oil available for supply in the next 3 or 4 months. The oil futures markets react quickly to expectations about differences in oil supply.

4. FUTURES MARKETS, OIL PRICE VOLATILITY, AND STOCK MARKETS

Oil is a commodity that is traded in both the spot and the futures markets. The spot market reflects trading for oil delivered today. The futures market reflects trading for oil delivered at some specified time in the future. The futures market provides an essential ingredient to risk management. In the oil market, futures prices tend to be followed more closely than spot prices because futures prices are not as affected by the short-term supply disruptions that affect spot prices. Futures markets for petroleum products are located in New York, London, and Singapore. The three major energy futures exchanges are the New York Mercantile Exchange (NYMEX), the International Petroleum Exchange (IPE), and the Singapore International Monetary Exchange (SIMEX). The NYMEX launched the futures contract on heating oil No. 2 in November 1978. In April 1983, NYMEX introduced the popular futures contract on crude oil. A futures contract on unleaded gasoline followed a few years later. The natural gas contract has only traded since April 1990. The NYMEX natural gas contract size is 10,000 million British thermal units of natural gas deliverable to Henry Hub, Louisiana.

The deregulation of the natural gas industry began in 1978 and was completed in November 1993. As deregulation evolved, the marketplace changed from one dominated by fixed long-term contracts (provided by the pipeline companies) to a spot market, followed by a forward market, and finally a futures market. The success of the natural gas futures contract in the 1990s enabled the natural gas industry to more efficiently manage risk through hedging strategies. In turn, the natural gas futures market has become the foundation for other types of derivatives and now forms the basis for financial engineering in the natural gas industry. Since its introduction in April 1990, the NYMEX natural gas contract has increased in importance as an effective financial instrument for natural gas price discovery.

In 1996, NYMEX launched two futures contracts on electricity. Trading in electricity futures is likely to increase in importance as North America experiments with energy market deregulation. NYMEX also trades options on futures for crude oil, New York Harbor unleaded gasoline, heating oil No. 2, natural gas, and electricity. Options on futures contracts are very useful for risk management.

The IPE trades two popular petroleum futures contracts in Brent crude oil and gasoil. Gasoil can be either heating oil or diesel fuel. Gasoil is the largest

single product market in Western Europe. The IPE also trades futures in natural gas. The SIMEX trades residual fuel oil and the market is relatively small compared to NYMEX and IPE.

The West Texas Intermediate (WTI) crude oil futures contract traded on NYMEX is the most closely watched oil price in the world. One contract of WTI crude oil is for 1000 barrels. The NYMEX crude oil market is the largest market (in terms of volume traded) outside of the financial futures market. The WTI crude oil futures market is clearly important to the United States. Moreover, the physical market underlying WTI crude oil is of particular importance to those interested in the demand and supply relationships that reflect the flow of oil from the Gulf of Mexico into the Chicago region. In a broader sense, the WTI crude oil futures contract has become a benchmark for the global oil industry. Reliable fundamental information on global demand and supply in the oil market is often difficult to obtain. Production estimates for some countries, for example, tend to be higher than actual production, and consumption data tend to be made available after some lag time. In contrast, reliable data on oil demand and supply in the United States are often easily accessible and very reliable. Given that United States oil consumption is approximately 25% of world consumption and the fact that the United States accounts for approximately 31% of global GDP, it is natural to use the U.S. data as a proxy for unobservable global fundamentals.

In some instances, futures prices can be used to provide a reasonably good forecast of future spot prices. In other cases, futures prices may provide poor forecasts of future spot prices. The key to determining when futures prices might reasonably be expected to forecast spot prices is to understand the role that storage plays in commodity markets.

Nonstorable commodities are those that are perishable and cannot be stored. Examples include agricultural products such as dairy products. For nonstorable commodities, futures prices reflect the relationship between demand and supply for commodities to be delivered in the future. Small changes in anticipated supply and demand conditions are immediately reflected in futures prices. In this case, futures prices can be used to forecast future spot prices.

Some commodities can be stored and have large inventories. For these commodities, futures prices are equal to the spot price plus the carrying costs associated with storage. In this case, futures prices provide little guidance to future spot prices. Examples include agricultural products such as corn and wheat and also metal products.

Some commodities, such as crude oil, can be characterized as storable commodities with modest inventory levels. In general, consumption is high relative to inventory levels. If futures prices are lower than spot prices, then the pricing structure is called backwardation. In this case, oil can be characterized as a nonstorable commodity. If oil supplies are expected to increase in the future, then the futures price would decrease relative to the spot price. This difference in price creates an apparent arbitrage possibility but low inventory levels prevent the actual collection of arbitrage profits. When commodity markets are in backwardation, futures prices are useful in predicting future spot prices.

If futures prices are higher than spot prices, then the pricing structure is called contango. In this case, oil can be characterized as a storable commodity with large inventory. If the supply of oil is expected to decrease in the future, then futures prices will be higher than spot prices. The difference between futures prices and spot prices reflects carrying costs associated with storage. When commodity markets are in contango, futures prices are of little use in predicting future spot prices.

Futures markets respond to news on four factors: demand and supply fundamentals, geopolitics, institutional arrangements, and the dynamics of the futures market. Whereas some of the spot markets for crude oil tend to be thinly traded, the futures markets (or markets for paper barrels) tend to be very liquid. Futures markets respond not just to facts about the demand and supply balance but also to rumors. In fact, some of these rumors can impact trader sentiment, which may be unrelated to the actual mechanics of the physical market for crude oil. Unanticipated changes in these factors, be it fact or rumor, can create volatility in oil futures prices.

Volatility in the oil futures market leads to uncertainty and this affects the decision making of consumers, producers, investors, and governments. Economic uncertainty creates situations in which planned spending on consumption and investment is postponed or canceled. This economic uncertainty affects the consumption, investment, and government spending categories of the expenditure side of a country's national accounts. Oil price uncertainty impacts the profits of firms, which in turn creates more uncertainty in the stock markets. Uncertainty in the stock markets usually translates into stock price declines, which further affect the economy either through reductions in consumer wealth or postponed investment by business. Volatility in oil prices makes short-term oil price forecasting very difficult. In the

long term, oil price fundamentals should dominate, making it easier, at least conceptually, to forecast the price of oil. Forecasting oil prices and oil price volatility, in general, is very difficult, but progress has been made in recent years by using new statistical techniques in modeling and simulation.

5. OIL PRICES AND CURRENCY EXCHANGE RATE RISK

There is evidence that currency exchange rates impact the movement of commodity prices. Since commodities are homogeneous and traded internationally, it is very likely that they are subject to the law of one price. This means that commodities have similar prices in each country's home currency. Thus, as the U.S. dollar weakens relative to other currencies (*ceteris paribus*), commodity consumers outside the United States should be willing to pay more dollars for commodity inputs priced in U.S. dollars. A strengthening U.S. dollar relative to other currencies (*ceteris paribus*) increases the price to commodity consumers outside the United States resulting in lower demand for the commodity. Most internationally traded commodities, including crude oil, are priced in U.S. dollars. Consequently, currency exchange rate movements may be an important stimulus for energy price changes. In particular, currency exchange rate movements impact crude oil prices and crude oil prices affect the prices of derivative products, such as heating oil and unleaded gasoline. Crude oil price changes also affect the price of natural gas. Thus, changes in crude oil prices create a ripple effect throughout the pricing structure of petroleum products. The dominance of the U.S. dollar may be challenged in the future as European countries make more use of the Euro currency in their business transactions. Already, some European countries prefer to price oil in Euros and measure oil in metric tonnes. Thus, there are times when the price of crude oil changes not so much from changes in the demand and supply balance in the global oil market or changes in OPEC pricing structure but instead from changes in currency exchange rates.

6. TECHNOLOGY STOCK PRICES AND ENERGY PRICES IN THE NEW ECONOMY

There are two views on the relationship between oil price movements and technology stock prices. The

first view is that oil price changes have very little impact on technology stock prices because technology companies engage in business activities that are not very energy intensive. The second view is that oil price changes do have an impact on technology stock prices because oil price increases fuel inflation and inflation leads to economic downturns. High-technology companies are very sensitive to the overall business cycle. Both views deserve further discussion.

The basis for the first view is the fact that technology companies produce products or services that are less energy intensive than products produced in other sectors of the economy. For example, it takes less energy per unit of output to produce telecommunications equipment than it does to produce steel beams. Manufacturing, in general, has also become less energy intensive. In the United States, it takes 45% less energy to produce one dollar of GDP today than it did in 1973. Automobiles built today are 50% more fuel efficient than automobiles built in the 1970s. The combination of less energy-intensive manufacturing and energy efficiency suggests that technology stock prices are relatively unaffected by oil price movements.

The second view is that oil price movements have a major impact on technology stock prices. Technology stocks, especially prices of Internet stocks, tend to be valued at very high price earnings multiples and stock price movements are linked to expected earnings. These stocks are highly vulnerable to market cycles. In particular, technology stocks, like other growth stocks, are very susceptible to inflation because inflation reduces corporate earnings. Furthermore, inflationary pressures tend to lead to higher interest rates and a slowing economy, which lessens the ability of growth companies to meet revenue and earnings targets, causing stock prices to decline.

Oil price increases are one of the most likely sources of inflation. Oil price increases usually impact the transportation sector first, with initial price hikes being absorbed by this sector. Persistent oil price increases lead to higher producer prices and eventually, after 6–9 months, inflation. Inflation, which erodes the real value of investments, leads to higher interest rates as contractionary monetary policy is used to bring economy-wide price increases down. In the high-technology sector, oil price increases can directly impact technology stock returns by raising the costs of the transportation component of e-business. When a customer places an order for a product online, the product still needs to be shipped from the warehouse to the customer's

doorstep. Higher interest rates also create competition for equity financing, and equity financing is more important than debt financing in the high-technology sector. In March 2000, for example, the debt equity ratio for NASDAQ listed stocks was 16.6%, whereas that for the rest of the U.S. stock market was 85.2%. Moreover, some large technology companies such as Cisco Systems have no debt. As interest rates rise, private investors may choose to buy fixed-income securities rather than invest in venture capital pools to finance the startup of new technology companies.

7. MODELING APPROACH AND QUANTITATIVE METHODS

There are three main ways to empirically investigate the interaction between stock prices and energy prices. The first approach uses arbitrage pricing theory (APT), the second approach uses a discounted cash flow model, and the third approach uses unrestricted vector autoregressions.

The APT approach uses an arbitrage pricing argument to relate priced risk factors to required rates of return in assets (securities):

$$R_t = \alpha + \sum_{j=1}^k \beta_j F_{jt} + \varepsilon_t, \quad (1)$$

where R_t is the expected excess return on an asset or stock market index, β_j are the factor loadings or systematic risks, and F_{jt} are the risk factors. The variable ε_t is a random error term. The APT assumes that the total variability of an asset's returns can be attributed to market-wide risk factors that affect all assets and other risks that are firm specific. Market-wide risk factors are known as systematic or non-diversifiable risks, and firm-specific risk factors are known as unsystematic or diversifiable risks. A widely diversified portfolio of equities can reduce or eliminate unsystematic risks, but systematic risk cannot be diversified. Thus, an investor must be compensated for systematic risk. In APT models, assets can have exposure to a number of different risks, including market risk, inflation risk, default risk, exchange rate risk, and oil price risk. Equation (1) can be estimated by ordinary least squares (OLS) regression using a time series data set on a single asset or as a cross section for one particular point in time with many assets. Model adequacy can be checked using regression diagnostic tests. More sophisticated models may involve many assets or classes of assets.

In this case, a system of equations estimation approach would be used. There is considerable evidence that asset prices display time-varying risk premiums and volatility clustering. Time-varying risk premiums can result from time variation in the risk sensitivities (factor loadings) or time variation in the market premiums (factors). Volatility clustering means that major stock price changes follow major stock price changes but the direction of the change is not known. As a result, time-varying autoregressive conditional heteroskedasticity models are often used in the estimation of Eq. (1).

The second approach uses a discounted cash flow model to test whether stock prices react rationally to changes in oil prices:

$$RS_t = E_{t-1}(RS_t) + (E_t - E_{t-1}) \sum_{j=0}^{\infty} \rho^j \Delta C_{t+j} - (E_t - E_{t-1}) \sum_{j=1}^{\infty} RS_{t+j} + \sum_{s=0}^k \theta_s OIL_{t-s} + \eta_t, \quad (2)$$

where RS_t is the log real return on a stock in period t , E_t is the mathematical expectation operator, C_t is the log of real cash flow in period t , ρ is a parameter close to but not equal to 1, OIL_t is the percentage change in oil prices, and η_t is a random error term. The model can be estimated by OLS. If the stock market is rational, then the coefficients on the OIL variables should be statistically insignificant from zero. At the economy-wide level, aggregate future cash flows can be approximated by future industrial production.

The third approach uses an unrestricted vector autoregression (VAR). This approach is intentionally unrestricted and useful for determining or confirming patterns between economic variables. Many interesting dynamic relationships can be captured using multiequation regression models such as VARs. Moreover, the VAR is very useful for performing policy simulations. The basic form of the VAR treats all variables symmetrically. No reference is made to distinguish which variables are independent and which variables are dependent. Consider a congruent statistical system of unrestricted forms represented by Eq. (3):

$$\mathbf{x}_t = \sum_{\tau=1}^p \Pi_{\tau} \mathbf{x}_{t-\tau} + \varepsilon_t, \quad \varepsilon_t \sim \text{iid}(0, \Omega), \quad t = 1, \dots, T, \quad (3)$$

where $\mathbf{x}_t$ is a $(n \times 1)$ vector of variables, and the error term ε_t is independent and identically distributed with mean zero and variance Ω . Economic theory

guides the choice of variables. Typical variables used in studying the relationship between stock prices and energy prices include industrial production, interest rates, stock prices, oil prices, consumer prices, and, possibly, exchange rates. The variables in the vector $\mathbf{x}_t$ might also be transformed into growth rates or may be estimates of volatility. In the latter case, the VAR would be used to investigate the dynamic relationship between oil price volatility and stock market volatility. Basic research questions include the following: What are the determinants of stock market volatility? Has stock market volatility increased over time? Has international financial integration led to faster transmission of volatility across international stock markets?

8. EMPIRICAL EVIDENCE ON THE RELATIONSHIP BETWEEN OIL PRICE MOVEMENTS AND STOCK PRICE MOVEMENTS

The empirical evidence on the relationship between oil price movements and stock price movements is mixed. Empirical models can be estimated using daily, monthly, quarterly, or annual data, and the analysis can be conducted at the firm, industry, or economy-wide level.

Empirical models using the APT model generally use monthly data to investigate the relationship between oil price movements and stock price movements across industry sectors or at the economy-wide level. Generally, the empirical results indicate that an oil price factor is an important driver behind stock price movements. At the economy-wide level, an oil price risk factor is important but it does not have the same impact on all countries' stock markets. Countries that are net importers (exporters) of oil and gas tend to have a negative (positive) correlation between oil price changes and stock returns. The Group of Seven (G7) economies (Canada, the United States, England, France, Germany, Italy, and Japan) is adversely affected by oil price increases. This effect is muted in Canada since it is a net exporter of energy products.

Industries are not all the same and different risk factors can have different effects across different industries. Empirical results suggest that factor sensitivities vary across industries. At the industry level, oil price risk factors have a positive and significant impact on oil company stock returns. For Canadian and U.S. oil and gas companies, market

betas are typically approximately 0.7 and oil price betas are typically approximately 0.3. For industries that are net users of oil and gas, such as airlines and trucking companies, oil betas have a significant negative impact on stock prices.

Empirical results from a discounted cash flow model to test whether stock prices react rationally to changes in oil prices are mixed. The reaction of Canadian and U.S. stock markets to oil price shocks can be completely accounted for by the impact of oil price shocks on real cash flows. The results for the United Kingdom and Japan are not very strong. For these countries, their respective stock markets seem to overreact to oil price shocks.

The VAR approach has been widely used in studies investigating the relationship between oil price movements and stock price movements. This approach, which is generally applied at the economy-wide level using monthly or quarterly data, has provided evidence across a variety of developed and emerging economies showing that oil price changes do impact stock returns. One of the more interesting discoveries from these studies is that oil price changes have asymmetric effects on economies and the stock market. Positive oil price changes exert a much larger impact on economies and stock markets than do negative oil price shocks. This is because positive oil price shocks are associated with an increase in uncertainty. In periods of high uncertainty, firms postpone their hiring and investment decisions. This slows GDP growth. Negative oil price shocks tend to be viewed as temporary and unlikely to affect the investment and hiring decisions of firms. Any benefits of lower oil prices are usually seen in increased profits for energy-consuming companies.

The VAR approach applied at the industry or firm level using daily data to investigate the lead-lag relationship between oil futures prices and stock prices during the day has provided little evidence of a relationship between broad-based stock prices and oil prices. This is probably because oil price changes take several months to impact the non-oil-related sectors of an economy. As expected, there is strong evidence to support the hypothesis that daily oil company stock prices respond quickly to daily oil price movements.

The VAR approach is also useful for studying the relationship between stock market volatility and the volatility of oil prices. Empirical results indicate that the conditional volatilities of oil prices, interest rates, and U.S. inflation rates each have a significant impact on the conditional volatility of equity markets.

9. ENERGY PRICE RISK MANAGEMENT

Fluctuating energy prices create uncertainty for both consumers and producers of energy products. Energy consumers prefer low and stable energy prices, whereas energy producers prefer higher energy prices. This situation creates a demand for risk management practices. There are a number of risk management tools, including forward contracts, futures contracts, options on futures contracts, and other derivative products (some traded and others not traded).

A typical example is an electricity power generator that needs to buy a known amount of natural gas 6 months in the future. The power generator is a heavy user of energy products and is concerned that natural gas prices will increase in the future. Rising natural gas prices will increase costs and lower profits, possibly resulting in declining earnings, a cut in dividends, and declining share prices. A useful risk management strategy is to employ a hedging strategy. The power generator could buy a forward contract for delivery of the natural gas in 6 months. In this case, the power generator buys the required amount of natural gas from a seller at a price negotiated today. This locks in the price at today's forward prices. The power generator could also use the futures markets for natural gas to buy the appropriate number of natural gas futures contracts on the NYMEX for delivery of natural gas 6 months in the future. Provided that the price of natural gas increases enough in the 6-month period, the forward or futures contracts will have provided the power generator with an effective hedge against rising natural gas prices. Hedging strategies that use forward or futures contracts eliminate the upside risk by locking in the price for natural gas. This eliminates the company's exposure to rising natural gas prices but, unfortunately, also gives away all or part of the company's exposure to declining natural gas prices.

Buying options on natural gas futures contracts is another form of risk management available to the power generator. In this case, the power generator pays the premium to purchase the options. The options premium is a form of insurance against rising natural gas prices. The power generator has the right but not the obligation to purchase the underlying futures contracts. The options will be exercised if natural gas spot prices rise high enough relative to the futures prices specified in the options contract.

If spot natural gas prices decline, then the options expire worthless. The company loses the premium paid on the options but can still buy natural gas in the spot market at a lower price than would have been available under the forward or futures contract.

SEE ALSO THE FOLLOWING ARTICLES

Business Cycles and Energy Prices • Derivatives, Energy • Economics of Energy Demand • Economics of Energy Supply • Energy Futures and Options • Inflation and Energy Prices • Innovation and Energy Prices • Markets for Petroleum • Oil Price Volatility • Prices of Energy, History of • Trade in Energy and Energy Services

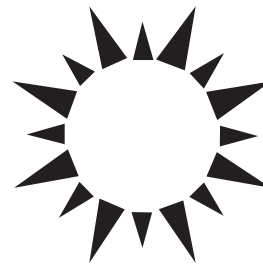
Further Reading

- BP (2001). Statistical review of world energy. www.bp.com.
- Engle, R. F. (1982). Autoregressive conditional heteroskedasticity with estimates of the variance of UK inflation. *Econometrica* 50, 987–1008.
- Ferson, W., and Harvey, C. R. (1994). Sources of risk and expected returns in global equity markets. *J. Banking Finance* 18, 775–803.
- Fusaro, P. C. (1998). “Energy Risk Management: Hedging Strategies and Instruments for the International Energy Markets.” McGraw-Hill, New York.
- Hamilton, J. D. (1983). Oil and the macroeconomy since World War II. *J. Political Econ.* 91(2), 228–248.
- Huang, R. D., Masulis, R. W., and Stoll, H. R. (1996). Energy shocks and financial markets. *J. Futures Markets* 16(1), 1–27.
- Jones, C. M., and Kaul, G. (1996). Oil and the stock markets. *J. Finance* 51(2), 463–491.
- Sadorsky, P. (1999). Oil price shocks and stock market activity. *Energy Econ.* 21(5), 449–469.
- Yergin, D. H. (1991). “The Prize: The Epic Quest for Oil, Money and Power.” Simon & Schuster, New York.

Storage of Energy, Overview

MARCO SEMADENI

Swiss Federal Institutes of Technology
Zurich, Switzerland



-
1. Introduction
 2. Electricity Storage
 3. Thermal Storage Systems
 4. Chemical Storage Systems
 5. Conclusion

Glossary

aquifer thermal energy storage (ATES) ATES systems store thermal energy in aquifers (natural groundwater basins or layers at different depth). The system consists of two groups of wells with hydraulic coupling, one cold and the other warm. In winter, ambient air or surface water can be preheated before use by a heat exchanger circulating water from the warm well to the cold. In the summer, excess heat, solar energy, or waste heat can be stored by reversing the process.

compressed air energy storage (CAES) Off-peak electrical energy is used to compress air into underground storage reservoirs until times of peak or intermediate electricity demand. To generate power, the compressed air is heated and then passed through turbines.

duct/borehole thermal energy storage (DTES) DTES systems store thermal energy using vertical ground heat exchangers in drilling boreholes. The thermal energy is transferred via plastic tubes in boreholes typically of 50–200 m depth. A heat carrier fluid is circulated through the tubes to transfer the energy between source and sink.

final storage The storage of final energy such as heat, which is used by the customer to sustain continuous energy services.

grid-level power storage Consists of small and high-power storage systems with the capability to improve reliability, flexibility, and quality of electricity services.

latent heat storage Storage of the energy in special materials such as phase change materials (PCMs). The energy required to change the state of a unit mass of material from solid to liquid or liquid to gas is stored without changes in temperature. This energy is then released when the material reverts from gas to liquid or from liquid to solid.

mechanical storage Energy storage by mechanical processes such as rotation.

secondary storage Storage of secondary energy.

thermal energy storage Storage of heat or cold in different medias, such as PCMs, thermo-oils, or water kept in insulated tanks, caverns, pipes, or underground. Heat storage is a secondary storage type (tertiary, if heat is generated by electricity).

underground thermal energy storage (UTES) Storage of thermal energy underground using either underground storage tubes, aquifers with reduced storage volume needs, or man-made cavern storage systems with sealed storage volumes.

uninterruptible power supply (UPS) systems Combination of high-power, high-energy-capacity storage components that work together to ascertain supply and quality for a specific application.

Energy supply is an intricate task that provides reliable energy service to consumers throughout the year. Import dependencies, seasonal differences in energy supply and use, and daily fluctuations in consumption require sophisticated management of energy resources and conversion, or energy distribution and resource intermittency, in order to guarantee continuous energy services throughout all sectors. Energy storage thus plays a critical role.

Energy is mostly stored in between conversion steps from primary to secondary energy and secondary to final energy. Often, rechargeable systems are used to refill the storage capacity when energy demand is low and energy services are not needed. Primary storage, such as large crude oil and natural gas storage tanks, is essential for the proper functioning of a country's energy system. Water stored in reservoirs behind dams is valuable for selling hydropower electricity at the right time or during seasons of increased demand. Secondary or final storage systems, such as in tanks or in batteries, are crucial for emergency situations, uninterrupted industrial production, long-distance mobility, or to secure energy services at home.

Storage systems are engineered to hold adequate amounts of mechanical, thermophysical, electrochemical, or chemical energy for prolonged periods of time. Energy storage systems should be quickly chargeable and should have a large energy storage capacity, but at the same time they should also have high rates of recovery and high yields of energy regain.

Final energy in factories or households is often stored in tanks as chemical energy in the form of heating oil or natural gas or as thermophysical energy in the form of steam, hot or cold water, or thermo-oils. For special applications or for safety reasons, energy may be stored electrochemically in batteries or physically in the form of pressurized air. Other storage systems are related to electricity and apply mechanical storage in the form of spinning turbines or flywheels, physical storage in the form of water in reservoirs in highland terrains, or electrostatic storage in supercapacitors.

1. INTRODUCTION

Several different types of energy storage systems are used for primary, secondary, and final energy. Primary energy is predominantly stored chemically in oil or gas tanks or in piles of specific energy carriers, such as biomass or coal. Secondary and final storage involve silos or piles of briquettes or biomass chips or storage tanks for refined petroleum oils, fuels, and gases. The conversion of these forms of energy into useful end-user energy provides for immediate energy services. If the demand or system status changes, useful energy can also be stored using mechanical or physical storage systems, thermophysical or thermochemical systems, electrochemical and electrostatic systems, or chemical storage systems.

For proper functioning of energy supply chains, storage concepts must be well developed to appropriately incorporate the storage technology into the energy system. Imbalances in a domestic energy system for heat, gas, and electricity can result in a country's increasing dependence on foreign gas and oil companies and expensive transboundary high-voltage electricity transmission lines. The development of inexpensive storage options may enable a country to cope with imbalances between demand and capacity and variable prices in energy markets. Uninterrupted availability of affordable clean energy products for both private and industrial consumers is important for society.

Energy systems rapidly change under the influence of energy policies. The ongoing liberalization or deregulation of energy markets may make small-scale and large-scale storage capabilities increasingly beneficial because of emerging competition in the energy market. However, increasing oil prices indicate structural weaknesses, with the increasing energy dependence on oil playing a major role in fuel consumption. Strategic petroleum reserves are still an important regulatory aspect of energy supply systems and compulsory storage ratios are negotiated between importing companies and the government. Adequate storage facilities must be available not only for crude oil but also for natural gas, which is expected to largely substitute oil in the near future.

Electricity and renewable energy technologies, increase variations in the grid and must be integrated properly into the existing electricity system. This involves local storage of electricity, which may require, among other things, installation of grid-level power storage capabilities.

Industrial and commercial customers increasingly demand power quality and reliability for functioning microprocessor systems. The spread of information technology brings new requirements for the transmission and distribution of electricity.

Uncertainties in grid reliabilities shift the emphasis in energy storage to guarantee energy service from producers to distributors. Distributed power generation and storage interconnected with the distribution grid can provide the needed power quality. In addition to energy supply reliability, incentives for energy storage include the moderation of increasingly volatile energy prices. Real-time pricing in liberalized energy markets is expected to result in additional local storage within buildings and "intelligent" household appliances.

To achieve full potential, both independent and grid-based power distribution systems rely on the availability of different storage technologies. These include safe on-vehicle storage and infrastructure components to ensure fuel availability at filling stations. Other examples are hybrid systems that combine microturbines, photovoltaic panels, or fuel cells for diverse industrial, commercial, or residential applications. These custom power systems must be electronically controlled within a complex grid containing multiple generating units. They require additional storage systems, such as batteries, superconducting magnetic energy storage (SMES), and flywheel energy storage, for power quality and reliability. In both small uninterruptible power supply (UPS) systems for personal computers and

large pumped storage projects, energy storage will increase system reliability. At utility sites, however, storage capabilities that provide large-scale instantaneous response are required. In the future, underground storage technologies including SMES may be possible. Innovations for storing hydrogen, recharging batteries or fuel cells, and facilitating high-temperature superconductivity have great potential to satisfy future heat and electricity storage requirements. In addition, new regulatory standards for efficiency and environmental compatibility require faster technical or technological innovations, such as heat-, corrosion-, abrasion-, and impact-resistant materials or thermal and chemical functional materials for thermal storage, insulation, and cooling media.

In the following sections, the basics of different types of energy storage systems are discussed for electricity, thermal, and chemical storage.

2. ELECTRICITY STORAGE

Apart from the use of secondary batteries, electricity storage is predominantly in the form of mechanical energy. Large-scale electricity storage systems are suitable to balance differences between power supply and demand. Power must be deliverable in time frames of hours and days or months to balance seasonal changes. Large-scale electricity storage systems supply peak-load power using previously stored baseload electricity or make it possible to integrate renewable energy technologies characterized by intermittent resource availability (e.g., solar and wind). In addition, they can supply regulating power for proper transmission and distribution of electricity through the grid. As part of the ancillary services, minimal power must be available within seconds or minutes. Sometimes, power levels must be sustained for hours. The following storage technologies are often used for large-scale electricity storage systems: pumped storage hydropower reservoirs, compressed air energy storage, and large-scale secondary batteries and large fuel cells.

Smaller scale applications of electricity storage provide power for mobile systems or upgrade power quality and reliability (including emergency supply). Electricity storage for these applications has varied capacity and timescale characteristics. The following technologies are applicable: batteries and fuel cells, SMES, supercapacitors and ultracapacitors, and flywheel energy storage.

Batteries are often used in mobile applications and systems in which reliability with low maintenance is important. Their release of power is instantaneous. Batteries can provide relatively prolonged power compared to capacitor and flywheel energy storage, which have different discharge characteristics, releasing high power in a short time frame. Capacitors and flywheels are suitable for improving power quality, integration into emergency backup systems, and peaking in mobile systems. Although in the testing phase, fuel cells are suitable for providing power over long periods of time and for continuous operation at various power ranges.

The capital costs of both power and energy are important in evaluating the economics of storage systems. Typically, the capital cost component of power is related to the output capacity, whereas the cost of energy is related to the storage capacity of the system (Table I).

2.1 Mechanical Systems

2.1.1 Hydropower Storage

An important application of hydropower storage plants is to balance seasonal differences in electricity demand. The reservoir stores excess water from water-rich, lower demand seasons and generates electricity during water-poor, higher demand periods. Pumped-storage hydropower plants follow a different strategy. Extra water is pumped back up to the reservoir using baseload electricity when electricity demand is low, usually at night, and is used for electricity generation during peak-load periods. The

TABLE I

Estimates of Power Capacity Cost and Energy Capacity Cost for Different Storage Systems^a

Electricity storage system	Power capacity costs (\$/kW/a)	Energy capacity costs (\$/kWh/a)
Compressed air energy storage in tanks	120	100–1500
Underground compressed air energy storage	90	5–100
Large-scale batteries	70	10–700
Pumped storage hydropower	50	3–90
Superconducting magnetic energy storage	40	9–150
Flywheel energy storage	30	80–1200

Annual costs assume a discount rate of 9% and a 10-year life cycle.

^aSource. EESAT (1998).

increase in electricity generation from renewable energy sources, such as wind and solar, requires additional storage due to its intermittent generation.

2.1.1.1 Hydropower Storage Plants Hydropower storage plants accumulate the natural inflow of water into reservoirs (i.e., dammed lakes) in the upper reaches of a river where steep inclines favor the utilization of the water heads between the reservoir intake and the powerhouse to generate electricity. This efficient storage of potential energy allows hydropower storage schemes a broader range of energy benefits than pure run-of-river schemes. Reservoirs at the upper watershed regulate the river downstream, which typically flows more evenly throughout the year, and the run-of-river power generated downstream utilizes part of the same water used upstream for the hydropower storage plant. Reservoirs can provide a seasonal, annual, or multi-annual storage capacity. Often several hydropower plants form larger hydropower schemes to follow demand along the load curves. Hydropower produces electricity preferentially during periods of higher load. The use of large storage volumes is often controversial because of considerable environmental impact, including the issue of pumping water from distant catchment areas to enhance inflow to the reservoir.

2.1.1.2 Pumped Storage Hydropower The most widespread large-scale electricity storage technology is pumped storage hydropower. A pumped storage plant uses two reservoirs—an upper storage basin providing the head to drive the hydropower turbines and another to collect water back into the upper basin using surplus baseload electricity during off-peak hours. The stored energy is only utilized for electricity generation during daily peak load. By International Energy Agency definition, a pure pumped storage plant uses less than 5% of inflow from upper watersheds and is therefore an energy system storage component, not a renewable energy source. Using “green electricity” (e.g., from surplus wind power) for pumping and using the additional storage volume for grid regulation and balancing are interesting possibilities for a future sustainable energy system.

Hydraulic turbines have a very high efficiency of more than 95%, whereas pumps are less efficient. The result is a maximum efficiency of approximately 85% for pumped storage hydropower plants. The effective efficiency is determined by the roundtrip cycle efficiency, which calculates the cycle with the

turbine operated both as a generator and as a pump. Designs of pumped storage are optimized for overall efficiency and dynamic response. The effective efficiency is 65–70%, and for each kilowatt-hour (kWh) of electricity produced, approximately 1.4 kWh of energy is necessary for pumping, assuming the same architecture for both operation modes. Pumped storage is one of the most efficient ways to store electricity.

Economic performance is highly dependent on the price of the cheapest electricity produced by other plants and the variation in daily load. A pumped storage project will provide the most efficient and cheapest operation when it can provide a high water head and volume between its two reservoirs. This will allow the greatest amount of energy to be stored in the smallest volume of water, resulting in smaller pumps and turbine, thus reducing capital costs.

Appropriate location, equipment, and construction are particularly important in view of the high costs associated with pumped storage systems. In general, pumped storage sites with high heads and short water passages are more desirable (i.e., cost-effective sites with higher water head ranges between 300 and 800 m with relatively steep topography). Shorter water passages reduce the need for surge tanks to control transient flow conditions. With higher heads, smaller volumes of water provide the same level of energy storage, and smaller size water passages can be used for the same level of power generation.

The global pumped storage hydropower capacity is estimated to be 82,800 MW. Capital costs are likely to be in line with those for a conventional project (i.e., \$1000–3000/kW). The payback time depends on the difference in value of peak and off-peak power. The availability of suitable sites in the future will limit growth in pumped hydrocapacity, with initial capital costs for construction increasing due to environmental obligations. In the longer term, the role of pumped storage plants is likely to increase to cope with intermittency of renewables.

2.1.2 Flywheels

Flywheel energy storage systems provide highly reliable, high-quality, uninterruptible electric power for communications networks, computers, the Internet, industrial manufacturing, commercial facilities, and distributed generation applications. Very heavy flywheel systems have been used for many years in electricity plants and smaller test systems have been used for mobile units.

The charging of a flywheel occurs when heavy symmetrical circumferential masses are rotated. The masses can rotate at approximately 50,000 revolutions per minute almost without friction around a titanium axle using magnetic bearings. The rotating masses are used in conjunction with a dynamo to generate electricity on demand.

The higher the angular velocity of rotation (ω radians per second) and the further away from the center the flywheel masses are placed, the more energy is stored. The relationship is $1/2I\omega^2$, where I , the moment of inertia of the flywheel, is the product of the total masses M and the square of their mean distance r from the center. Since the centrifugal forces are equal to $M\omega^2r = I\omega^2/r$, for a given flywheel diameter and thickness, higher stored energy requires stronger flywheel materials to withstand the rotational forces. For materials such as fused silica, the specific energy stored in watt-hours per kilogram (Wh/kg) is more than 20 times better than that of steel-based flywheels and approximately 40 times better than that of lead-acid electric vehicle batteries (Table II).

Modern flywheel energy storage systems make use of advanced composite materials and state-of-the-art active bearing technologies. Single flywheel units are typically 5–100 kWh and can be cascaded to achieve utility-scale power and energy ratings with much larger systems being developed in various research laboratories. A 2-kWh composite flywheel system has been introduced and flywheels with an increased energy of up to 4 kWh are available. Plans include production of an 8-kWh model, which is to be followed by a 20-kWh unit. Medium-scale storage (0.1–10.0 MWh) has some unsolved technological problems. Experiments pairing microturbines with a fuel cell or flywheel system as a complete hybrid backup power system are under way.

TABLE II
Energy Densities of Storage Media^a

Media	Wh/kg ^b
Hydrogen	38,000
Gasoline	13,000
Flywheel, fused silica	870
Flywheel, carbon fiber	215
Flywheel, steel	48
Lead-acid battery	25

^a Source: Ristinen and Kraushaar (1999).

^b Weight of motor and generator and conversion efficiency not included.

Ceramic superconducting materials may also be applied—for example, a flywheel running on frictionless bearings by means of magnetic fields generated by the superconducting materials.

2.1.2.1 Flywheels for Vehicles Electric buses or trains could use rapidly rotating flywheels for energy storage instead of secondary batteries. Urban transit buses and military vehicles have used stored inertial energy since the 1970s. Some impractical aspects, including a short range of only approximately 10 km, have limited their applicability.

Modern flywheels can store much more electricity than contemporary metal hydride or lead-acid batteries of similar weight or volume. Flywheels are environmentally friendly and have a much longer lifetime than electrochemical systems. High-performance materials enable the design of electric vehicles with flywheel systems comparable to petroleum-powered vehicles. Similar acceleration, speed, and range per unit weight of energy stored, including regenerative braking, are achieved. Hybrid vehicles using flywheel energy storage performed similarly to electric vehicles with supercapacitor or battery systems. One goal is to develop a zero-emission vehicle combining hydrogen lean-burn spark-ignited engines with high fuel economy or hydrogen fuel cells, with both incorporating flywheels or supercapacitors as acceleration and braking energy recovery systems.

2.1.3 Compressed Air

In compressed air energy storage (CAES), the release of stored pressurized air is used to generate electricity, most efficiently in conjunction with a gas turbine. In a CAES plant, compressed air is used to drive the compressor of the gas turbine, which uses 50–60% of the total energy consumed by the gas turbine system. The most important part of the CAES plant is the storage facility for compressed air. Usually, a man-made rock, salt, or porous rock cavern, created either by water-bearing aquifers or as a result of oil and gas extraction, is used. In contrast to the other systems, aquifers can be very attractive because the compressed air will displace the water, setting up a constant pressure storage system.

CAES plants show very high availability and reliability. CAES technology in the range of 25–290 MW has been achieved throughout the industrialized world, and there are proposed projects for an installed power of up to 1050 MW. The estimated annual investment costs for a CAES are \$90–120/kW/year

(i.e., uniform cash flow), depending on air storage type.

The development of large-scale CAES is limited by the availability of suitable sites, with current research focusing on the development of systems with man-made storage tanks.

2.2 Electrochemical Systems

Electrochemical storage is characterized as the ability to convert chemical binding energy directly to electricity. The process can be reversed in secondary (rechargeable) batteries or accumulators to recharge the storage media. Only recently have large-scale applications of battery storage become of interest. Utility electricity storage requires a large battery system that can be repeatedly charged and discharged. Battery systems can release power virtually instantaneously, significantly improving the stability of an electricity network. Because of slow self-discharge and the aging processes, battery cells exhibit a decreasing capacity and recharging ability, which results in electricity storage for limited periods of time only. The potential advantages of battery systems are quietness, freedom from maintenance, reduction of gaseous emissions, and their independence of the primary energy carrier. Electric batteries also store direct current (DC) and can be easily integrated into domestic DC networks. An important system parameter is the charging (or load) factor, which relates to the necessary quantity of electricity in Ampere-hours for the batteries to become fully charged to the corresponding quantity released in the previous cycle.

The secondary battery is well suited for applications involving instant, relatively low-voltage power. Battery operation must be designed to be application specific and must consider factors influencing cost, which often limit energy and power per unit weight and volume (energy and power density, specific energy, and specific power) or durability (cycle lifetime and maintenance). Costs and performance of larger systems must be competitive with those of common mechanical storage systems (Table I).

2.2.1 Energy Storage in Batteries

The performance of a battery is characterized by thermodynamic and kinetic parameters that depend on the materials and media used and on battery design. Thermodynamically, large chemical-binding energy differences between the reaction partners, represented by a difference in Gibbs reaction energy, appear as cell voltage. To achieve a high specific energy (Wh/kg), the electrochemical reaction part-

ners or electrodes should be as light as possible. Conventional lead-(sulfuric) acid batteries store approximately 200–500 times less enthalpy than petroleum fuel. Significant improvements have been achieved with nickel-metal hydride and lithium-ion batteries compared to older lead and nickel-cadmium batteries (Table III). The main requirements are improved lifetimes and replacement of heavy metals such as lead with much lighter and nontoxic materials. In addition to the energy storage capacity of a cell, the speed and quality of recharge are often important for efficient and attractive applications of battery systems. The time required to recharge a battery is strongly dependent on operating temperatures. Based on the solubility of reaction products or possible side reactions inside the cell, the recharging process can be inhibited or stopped completely after a few recharge cycles.

2.2.1.1 High-Energy Density Batteries Electrochemical power sources can be classified as accumulators (secondary cells), which store all corresponding reactive components inside, or as fuel cells with reactants stored outside that are fed to the cell stack when power is needed. The systems are also distinguished according to their operating temperature and electrolyte. The most important high-energy density battery systems are summarized in Table IV.

High-temperature secondary batteries and fuel cells compete to claim the electric vehicle markets. In recent years, the fuel cell has received a great deal of attention since it offers good performance in lower temperature operation and allows greatly improved electric vehicle ranges.

Regenerative fuel cells can be used to store electricity in chemical form and convert the stored energy back into electricity when needed. A regenerative fuel cell plant could store more than 100 MWh of energy and provide power for hours to several thousand homes. The development of regenerative fuel cells faces many challenges, particularly regarding pole reversal and recharging cycles. A new regenerative fuel cell design, the redox-flow battery, stores electricity in liquid (dissolved) form, allowing the maximum storage capacity to be easily increased by adjusting the electrolyte volume. The energy density is 20–30 Wh/liter, which is approximately 30% of a normal lead-acid battery. Redox-flow batteries are relatively large and heavy, which makes them suitable only for stationary applications. A large-scale application with a capacity of 15 MW/120 MWh has been built based on the Regenesys

TABLE III
Battery Systems Used in Applications^a

System	Electrolyte	Type	Storage capacity (Wh/kg) ^b	Application
Lead–acid	Aqueous electrolyte (sulfuric acid), mostly immobilized	Often sealed (valve regulated), works at low temperatures, rechargeable	20–50	Starter battery for cars, batteries for forklifts, electric vehicles, submarines, and ships (also unsealed non-maintenance-free)
Nickel–hydroxide cadmium	Aqueous electrolyte, mostly immobilized	Sealed (valve regulated) rechargeable	20–55	Battery system mostly for small equipment and some for propulsion, on-board electronics, and on-site applications
Nickel–hydroxide hydrogen gas	Aqueous electrolyte, immobilized	Sealed, pressurized, rechargeable	50–60	Battery system for many satellites
Nickel–hydroxide metal hydrides ^c	Aqueous electrolyte, immobilized	Sealed (valve regulated), relatively high load capacity, rechargeable	50–80	Can replace nickel cadmium batteries
Sodium sulfur ^d	Solid electrolyte (β -alumina) at temperatures of 300–400°C	High-temperature system, rechargeable	90–120	Electric vehicles, trams, buses
Sodium nickel–chloride ^d	Solid electrolyte (β -alumina) at temperatures of 300–400°C	High-temperature system, rechargeable	90–100	Electric vehicles, trams, buses
Lithium–ion polymer, lithium cobalt oxide, lithium manganese oxide, iron, or aluminum disulfide	Organic electrolyte at room temperature	Relatively high voltage (4V), relatively high load capacity, rechargeable	~100	Battery system for small applications, tests for some electric vehicles
Zinc bromine ^e	Aqueous electrolyte	Rechargeable	65–70	Tests for some electric vehicles
Zinc air (oxygen from air) ^e	Alkaline or neutral electrolyte	Mechanically rechargeable (i.e., exchange of zinc electrode after a few cycles)	~170	Tests for some electric vehicles

^a Source. Kiehne, H. A., Berndt, D., Fischer, W., Franke, H., König, W., Köthe, K. H., Preuss, P., Sassmannshausen, G., Stahl, U.-C., Wehrle, E., Will, G., and Williams, H. (2000). “*Batterien: Grundlagen und Theorie aktueller technischer Stand und Entwicklungstendenzen*,” Expert Verlag, Renningen.

^b When applied.

^c Some nickel alkaline electrolyte batteries also work at low temperatures.

^d Sodium–sulfur and sodium nickel–chloride have been abandoned for traction applications.

^e Different battery system architecture (i.e., reactive substrates stored outside the cell).

regenerative fuel cell employing the bromide/tribromide and the polysulfide/sulfide couples in aqueous solution. Technical and economic data, according to National Power PLC, are given in Table V.

2.2.1.2 Battery Systems for Different Applications Although the end use of electricity is efficient when compared to the low energy conversion efficiency of gasoline, batteries still severely limit

electric vehicle range. No battery has the energy density of gasoline, and battery recharging takes longer than refueling. Nonetheless, by providing a low-emission, low-noise, and clean mobility environment, battery-powered automotive systems for indoor use are increasingly being applied in organized transport systems in automated industries.

Electric buses for public transport and electric cars are becoming more popular for environmental

TABLE IV

Properties of Secondary Batteries and Fuel Cells^a

Class	System	Type	Operating temperature (°C)	Electrolyte	Energy density (Wh/kg) ^b	Power density (W/kg) ^c	Lifetime charging cycles/years charged
Accumulators	Lead–lead oxide system	Pb/PbO ₂	<45	H ₂ SO ₄	161/35	80/160	750/>3
	Lithium–metal sulfide (oxide) system ^d	Li/TiS ₂	Ambient	Polymer	473/100	35/400	300/1
	Lithium–iron sulfide system	Li/FeS ₂	400	Salt melt	650/180	30/400	1000/0.9
	Zinc–bromine system	Zn/Br ₂	60	ZnBr ₂ + H ₂ O	430/55	70/80	400/2
	Sodium–sulfur system	Na/S	320	Ceramics	758/110	95/150	4000/4
	Sodium–nickel chloride system	Na/NiCl ₂	300–400	Ceramics	650/82	60/140	2000/—
					Fuel	Maximum power (W/cm ²)	Maximum operation time (h)
Fuel cells	Low-temperature fuel cells	H ₂ /air	<100	Polymer	H ₂ or MeOH	1500	1200 (1-kW module)
	Mid-temperature fuel cells	H ₂ /air	>100–200	H ₃ PO ₄	Natural gas	150	16000 (several MW)
	High-temperature fuel cells	H ₂ /air	600	Carbonate melt	Natural gas, gasification	150	5000 (3-kW module)
	High-temperature fuel cells	H ₂ /CO/air	1000	Ceramics	Natural gas, gasification	300	40000 (cell)

^aSource. Kiehne, H. A., Berndt, D., Fischer, W., Franke, H., König, W., Köthe, K. H., Preuss, P., Sassmannshausen, G., Stahl, U.-C., Wehrle, E., Will, G., and Willmes, H. (2000). “*Batterien: Grundlagen und Theorie aktueller technischer Stand und Entwicklungstendenzen.*” Expert Verlag, Renningen.

^bTheoretical/experimental: The theoretical value accounts only for the weights of the reaction substrates. The experimental value accounts for a 5-h discharging time, except for the sodium–sulfur system, which has 1 h.

^cContinuous operation/pulse operation (30 s).

^dOther combinations include MoS₂, NbSe, V₂O₅, and CO₂.

reasons. As Lindly and Haskew note, complete electrification of the vehicle fleet of gasoline-fueled, light-duty, internal combustion (IC) engines, including the complete fuel cycle and based on 1997 Alabama electricity generation mix, would provide a 44% decrease in nitrous oxide (NO_x) emissions, an 18% decrease in CO₂ emissions, but a 14% increase in sulfur dioxide (SO₂).

In contrast to fuel cells, research and development of electrical vehicles with lead–acid, nickel–cadmium, nickel–metal hydride, or molten salt

sodium β -alumina batteries has not achieved the required reduction in weight and volume to compete with conventional IC engines. An innovative concept is the hybrid vehicle with a small, efficient IC engine to drive the vehicle after the electric motor has accelerated the vehicle and to recharge its electrochemical system when idling. Recapture braking energy adds to the recharge, and with the reserve electric motor the idle time of the IC engine can be minimized, resulting in improved fuel efficiency.

Stationary fuel cells can also be used for public mobility as centralized electrical energy supply for trams and trolleys, underground (subway), light and conventional rail, and high-speed trains if combined with high-power storage systems such as superconducting magnetic energy systems (SMES) and ultracapacitors.

For large on-site immobile battery systems, mostly lead-acid- or nickel-based batteries are used in power plants and solar installations, signal transmission (telecommunications and radio) and UPS systems, backup systems, switchgear installations, annunciator and traffic lights, safety or emergency lighting, and air-conditioning installations. Because of the nature of the applications, high reliability is required. Less than a 0.25% failure rate has been observed with lead-acid and nickel-cadmium batteries in such systems.

TABLE V
Characteristics of Regenesys^a

Power	15-MW maximum
Storage capacity	120 MWh
Footprint	<5000 m ²
Storage tanks	Two tanks, 1800 m ³ each
Total installed costs	\$22 million
Lifetime	15 years
O&M costs	Relatively small (unmanned facility)
Storage time	10 h
Capital costs	\$185/kWh, \$50/kWh (mature technology)
Overall efficiency ^b	70%

^aSource. *Renewable Energy World* (2000, January).

^b<http://www.electrosynthesis.com/news/w8content.html>.

TABLE VI
System Power Range for PV Storage Applications^a

Power range	Typical application	Examples
μW	Integrated transistors or chips with minimal energy consumption	Solar watches and calculators
mW	Equipment and installation with low energy demand and only periodic use	Portable receivers/transmitter, automatic devices such as ticket or vending machines, fire or security alarm systems
W	Installations for communication and measurement, small consumer households, or businesses	Devices on buoys or for TV, radio, and meteorological stations, electricity supply on boats or in vacation homes, electricity supply for heat pumps
kW	Stand-alone grids with electric equipment and installations	Remote estates, military

^aSource. Kiehne, H. A., Berndt, D., Fischer, W., Franke, H., König, W., Köthe, K. H., Preuss, P., Sassmannshausen, G., Stahl, U.-C., Wehrle, E., Will, G., and Willmes, H. (2000). "Batterien: Grundlagen und Theorie aktueller technischer Stand und Entwicklungstendenzen." Expert Verlag, Renningen.

For telecommunication equipment, batteries need long operating times to balance electricity supply outages of several hours. For UPS systems, batteries must be able to instantaneously release very high currents over short time periods (minutes) until the backup system responds, which is comparable to starter batteries. Safety and durability are the most important criteria. In addition, sealed valve-regulated lead-acid batteries are used because of their low maintenance requirements.

Solar thermal and photovoltaic (PV) applications are always associated with battery subsystems to balance wide fluctuations of input and output currents. To determine the optimal system dimensions for the specific storage application, the average electricity consumption and the diurnal consumption profile should be known. Other important parameters include those of the PV cells, which must be adapted to the seasonal variation of solar radiation. Use of grid-dependent PV cells reduces the intermittent nature of solar energy because the grid and small-scale electricity storage devices are employed as a backup system. Other solutions to cope with the seasonal differences combine grid, PV, and liquid natural gas (LNG) burners into a coordinated hybrid system for buildings. Table VI lists storage applications for PV supply technologies.

2.2.1.3 Batteries and the Environment The increasing use of secondary batteries, especially lead-acid or nickel-cadmium batteries, means that hundreds of tons of toxic heavy metals are discharged yearly into the environment. Potential environmental impacts from their improper disposal have been recognized as a problem worldwide. In

Europe, the 1992 Government Action Plan for Waste and Recycling and the 1992 Waste Management Plan for Municipalities consolidated the framework of the Environmental Protection Act of 1994. In the United States, the 1996 Mercury-Containing and Rechargeable Battery Management Act (the battery act) promoted the collection and recycling or proper disposal of used nickel–cadmium and lead–acid batteries, clear labeling of batteries, and education of the public about recycling and disposal. Another goal is to phase out the use of mercury in batteries.

The use of nickel–cadmium batteries is proposed to be phased out of a wide range of household appliances or at least be restricted, for instance, to power tools requiring very rapid discharge. Another measure is to lower the threshold amounts of cadmium in batteries in a stepwise manner, which will ultimately require a change to alternative storage systems. Alternative secondary battery systems are lithium and nickel–metal hydride batteries.

The development of fuel cells for cars, trucks, and buses will reduce the environmental impact of mobility, even if natural gas supplies the fuel cells during transition stages. True “green mobility” may be achieved only if hydrogen supply and storage technology are developed using hydrogen generated from regenerative sources.

2.2.2 Ultra- or Supercapacitors

Capacitors are energy storage devices that use charge as a storage mechanism. A conventional capacitor with charge $+Q$ on one plate, $-Q$ on the other, and voltage V between them has a stored energy equal to $1/2 QV$. Almost all capacitors use a dielectric material between the plates, which optimizes charge distributions. Dielectric capacitors can provide power densities many times more than 1 kW/kg and have very long cycle lifetimes, whereas batteries provide less than 0.2 kW/kg and have fewer cycles (generally 100–1000). However, dielectric capacitors have energy densities less than 1 Wh/kg , whereas batteries can provide more than 100 Wh/kg . The applicability of dielectric capacitors is therefore clearly limited to specific tasks requiring high power at high voltages over very short time periods (e.g., for lasers). Superconducting inductive systems have similar limitations.

Supercapacitors (scaps) or electrochemical capacitors (ECs), which have 500,000 times larger capacitance than conventional capacitors, can provide more than 1 kW/kg power output combined with more than 5 Wh/kg specific energy and a high cyclicality of thousands of recharge cycles. Scaps, also

known as ultracapacitors or electrochemical double-layer capacitors, consist of carbon or other high-surface-area materials as a conductor and an electrolyte. Applications include acceleration power for electric and hybrid vehicles, electrical regenerative braking storage, power trains, starting power for fuel cells, pulse power for mobile and wireless telecommunications, and consumer and industrial electronic devices.

In contrast to batteries, scaps must be associated with a power supply system. They can be used as energy buffer banks to minimize the constraints of power supply. If coupled with batteries, they can reduce the peak power requirement, prolong battery lifetime, and reduce the energy requirement (or the size) of the battery. The batteries provide the energy, with scaps providing the instantaneous power needed. Depending on the configuration, scap systems can be used both as energy source and as buffer, or they can be applied as a load equalizer for elevators. Scaps work in PV-powered systems or stationary industrial power applications; as low-maintenance, solid-state alternatives to batteries for short-term bridging in UPS systems; or for peak-load buffering to increase efficiency and reduce the size and cost of stationary systems. As material costs decrease, low-cost scaps are expected to become the standard option of energy storage systems for consumer electronics, industrial, and transportation markets.

2.3 Electromagnetic Systems: Superconductivity

Superconductivity offers the ideal way of storing electric power directly. Theoretically, by using supercooled inductive coils, much higher storage capacities are achievable than in conventional capacitive or inductive systems.

Superconductors are being considered for SMES, in which electric energy is stored by circulating a current in a superconducting coil without resistive losses. Niobium–titanium alloys are used for storage at liquid helium temperatures (2–4 K). The use of superconducting materials with higher critical temperatures (e.g., 60–70 K) requires additional research and development. The efficiency of charging and discharging is very high because no energy conversion is involved; for this reason, SMES can respond rapidly, with a response that is limited by the time required for conversion between DC and AC power. Rapid response within seconds makes SMES attractive for load leveling and for frequency and voltage

control of the electricity grid while permitting a compact system design. Only SMES with relatively small storage capacities are commercially available. New SMES developments target storage capacities of approximately 100 kWh.

Large-scale applications depend on achieving high critical current densities (Jct) to be cost competitive with other technologies. High power and massive magnetic fields are involved, so careful consideration must be given to the consequences of an accidental loss of coolant; 10 MWh of stored energy released is equivalent to 1.18 tons of TNT.

SMES technology with 90% efficiency for storage of excess nighttime electrical energy is comparable to water-storage technology with efficiencies of 65–70%. Technological development, however, requires time and large monetary investments.

3. THERMAL STORAGE SYSTEMS

The energy content E of a thermal energy storage volume V is proportional to the temperature before (T_1) and after recharge (T_2). That is, $E = c_p(T_2 - T_1)V$, where c_p is the specific heat per unit volume—a storage media- or material-specific coefficient. Water has a high specific heat of 1 kcal/kg·°C compared to steel, which has a specific heat of 0.1 kcal/kg·°C. The transferred energy involved in temperature changes of materials is called sensible heat, whereas the energy involved in changing the phase of a material is called latent heat. Depending on restrictions of volume or pressure changes in gaseous or liquid storage media, the specific heat of constant volume C_v or of constant pressure C_p are important parameters in the design of efficient thermal energy storage systems.

Heat and cold storage systems are important for short-term and long-term (seasonal) storage. One of the benefits of short-term storage is to shift electric resistance heating to off-peak hours. Environmental benefits include the possibility of using renewable energy sources to increase the demand coverage. The benefit of long-term storage is that by balancing seasonal discrepancies in supply and demand, energy consumption can be significantly reduced.

The following systems exist for short-term storage:

- Water, which is the standard medium for heat storage
- Ice (phase change at 0°C, latent heat provides large energy density)

- Phase change materials (salt hydrates), which allow phase change at other temperatures besides 0°C are currently in the market introduction phase

- Sorption systems (e.g., H₂O/SiO₂ and silica gel), which are in the experimental stage, having higher energy densities are more suitable for long-term storage.

Other relevant parameters for thermal systems are energy losses as a percentage of the retrieved thermal energy, which depends on energy density, reflux temperatures, insulation quality, surface-to-volume ratio, and the duration of storage. The energy density (kWh/m³) depends on the temperature of the media and its specific heat capacity. The cost of thermal systems depends on volume and energy stored (Table VII).

Water as thermal storage media has a low energy density and consequently needs larger volumes, with correspondingly large heat losses. The surface-to-volume ratio should be as small as possible, especially for long-term storage. The recovery efficiency of heat from thermal storage depends on the reflux temperature of the heat extraction system. Improvements are possible using heat pumps, although this is more costly.

Technological advances in phase change materials (PCMs) can increase the thermal mass per unit floor area of buildings and improve the efficiency of thermal systems. Adsorption systems have great potential as seasonal storage applications of solar thermal energy for single homes, for which less cooling is required. The heat coverage with solar thermal energy could technically reach 90–100%.

Long-term storage is often combined with solar thermal systems. Soil-heat exchangers integrated in pile foundation, soil, or groundwater systems (aquifers) are used. The energy used for circulating the storage media water through the system is determined by the permeability of the underground

TABLE VII
Overview of Short-Term Storage^a

	Energy density (kWh/m ³)	Energy losses (estimate) (%)	Cost (<10 m ³) (\$/m ³)	Cost (>100 m ³) (\$/m ³)
Water	50–60	20	2000	700
Phase change material	50–120	12	2700	1300
Adsorption	200–250	5	6800	4500

^aSource. Collinson (2001).

material and the system's depth. The corresponding net thermal energy retrieved, the size, and the temperature levels of the system are the most important characteristics. The costs are strongly determined by the required pumping capacity (m^3/h) and power capacity (kW) and less by the size of the system (Table VIII).

Underground soil or groundwater systems are very productive for cooling, when improved building construction, (automated) shading systems, highly efficient lighting systems, and electric appliances reduce cooling demand to a minimum. Combined with energy conservation strategies, cooling systems can meet the entire cooling load in most office buildings. In winter, the same system can be used to preheat the cold air drawn in for fresh air supply in office buildings, reducing heating demand by approximately 30%. Seasonal underground storage systems can even provide coverage of up to 70% space cooling in industrial buildings with cooling capacities of $40\text{--}70\text{ W/m}^2$.

3.1 Water for Thermal Energy Storage

3.1.1 Thermal Storage Tanks

Centralized water thermal storage is by far the most common form of thermal energy storage. Usually, large hot-water storage tanks are buried underneath large infrastructure components such as athletic fields and parking garages. Conventionally welded steel tanks, reinforced concrete, or wire-wound concrete tank systems are used with capacities of 1 million gallons or more. Heat storage capacities range from 60 to 80 kWh/m^3 .

Low heat-conducting tanks ($<0.3\text{ W/m}^2\text{ K}$) have been developed in Germany using fiberglass composite materials. With volumes between 5 and 10 m^3 , these innovative tanks are applicable for buffer and

short-term storage in the range of hours to days. For storage in the range of weeks, volumes of $50\text{--}100\text{ m}^3$, and for long-term or seasonal storage, volumes of $500\text{--}6000\text{ m}^3$ can deliver thousands of ton-hours of hot water. The cost of large tanks (6000 m^3) is approximately $\$100/\text{m}^3$. Such tanks can also be used to store cold water.

Other thermal storage systems also used for cold-water storage are based on stratified storage underground tank systems combined with linear diffuser systems to reduce inflow turbulences. The different temperatures contained in a thermally stratified storage tank hold the density variation of warm and cold water. Thermally stratified storage tanks represent an effective technique that is widely used in energy conservation and load management applications. It is commonly applied in combination with solar energy or waste heat reuse systems. Thermal stratification improves the efficiency of storage tanks because heat at intermediate temperatures can be used to heat colder layers. System parameters include various inlet and outlet locations, velocities, temperatures, and heat loss. Heat conduction at tank walls and ambient heat loss through the cover strongly affect the stability of thermal stratification.

Long-term heat storage in combination with waste heat distribution, solar, or combined heat power (CHP) systems is also constructed using man-made water-permeable gravel pits insulated and covered with heat-damming materials, such as polypropylene foil, vacuum insulation foil, and glass granulates. The storage temperature is approximately 90°C . Heat storage capacity is $30\text{--}50\text{ kWh/m}^3$.

Short-term or buffer heat storage systems are ubiquitous for generation of hot water in heaters or boilers. The size of decentralized heat storage systems depends on the water boiler output power needed to satisfy hot water and heating requirements. A minimum of 25 liters storage volume per kilowatt boiler output is needed; for more comfortable operation, 70 liters/kW is recommended. Decentralized thermal storage heater/boiler systems mostly use electrical night currents for electrical resistance heating. An important aspect of the short-term storage concept is maintaining thermal stratification. The performance of storage-type domestic electrical water heaters is thus greatly improved and energy conservation optimized. Designs are dependent on draw-off rates (usually $5\text{--}10\text{ liters/min}$), with the placement of heating elements and residence time being important design parameters. Optimized ratios of tank size and draw-off rates together with better

TABLE VIII
Overview of Seasonal Storage^a

	Pumping capacity (m^3/h)	Power capacity (kW)	Pumping energy (% of useful energy)	Cost ($\$/\text{m}^3/\text{h}$)
Soil heat exchangers	0.5–50	5–450	5–8	1800–4100
Groundwater system	10–500	100–4500	5–8	1600–2700

^aSource. Collinson (2001).

thermal stratification inside the heater storage tank result in higher discharging efficiencies.

3.1.2 Large Underground Thermal Energy Storage Systems

Building, industrial, and agricultural sectors require much larger thermal storage dimensions. There are three forms of underground thermal energy storage systems (UTES):

- Thermal underground storage tubes (e.g., soil-heat exchanger), which require large storage volumes to improve storage efficiency, achieving 15–30 kWh/m³
- Aquifer storage systems with reduced storage volume but site-specific requirements, achieving 30–40 kWh/m³
- Man-made cavern storage systems with confined storage volumes, achieving 40–60 kWh/m³.

Early adoption of energy storage in standard project designs is essential to improve energy efficiency. In nonsealed UTES systems, heat is normally stored between 15 and 40°C, whereas cold storage areas are between 5 and 25°C. Depending on the size of the area used, heat/cold storage capacities of up to 5 or 6 GWh per year are achievable. Newer concepts of UTES at approximately 70–90°C provide more economically efficient and environmentally benign energy systems with reliable long-term operation. Other types of UTES systems are aquifer thermal energy storage (ATES) and duct/borehole thermal energy storage, which can store waste heat from municipal facilities in summer and provide domestic hot water and space heating in winter.

Chemical aspects of thermal energy storage in aquifers must be investigated to solve clogging and environmental heat exchange problems. For example, at high temperature ATES consisting of vertical wells with horizontal drains, major problems can occur at the heat exchange interface where large amounts of calcite form. Developing and testing site-specific groundwater treatment methods involving chemical, microbiological, and environmental effects ensure required environmental standards. Adequate scaling of ATES improves reliability and avoids or reduces potential ecological impacts.

The implementation of heat/cold storage applications as a standard design option improves effective storage control and operating strategies. This results in increased energy efficiency and cost-effectiveness in building and industrial fields. A combined heat/cold ATES installed in Berlin in a new government building complex uses two separate aquifers at

different levels, one at approximately 60 m for cold storage and the other at approximately 300 m for heat storage. The storage system is coupled with a heat pump and, in case of extra load in winter, with a CHP unit running on biofuels. At peak operation, this ATES system should provide 16-GWh heat and 3-GWh cooling capacity.

Man-made cavern storage systems are placed in leak-tight granite, gneiss, or hard sediment rock. Caverns can store heat at temperatures of approximately 90°C due to low heat-conducting properties. The cavern storage system of Lyckebo in Sweden has a storage volume of approximately 100,000 m³ and a heat storage capacity of 5.5 GWh.

3.2 Latent Heat/Phase Change Systems

Although it is inexpensive, readily available, and possesses excellent heat transfer characteristics, water is limited as a viable storage media because of the large volumes needed to achieve adequate heat storage capacities. When space is limited, phase change systems are applied to make use of latent heat.

The oldest PCM in thermal storage is ice. Ice is used in glycol static ice systems with phase change at 0°C and an energy density of approximately 90 Wh/kg for cooling. Glycol static ice systems are costlier than cold-water systems but require less space. Heat storage at various temperatures can also be designed using other PCMs (Table IX). The overall temperature required to melt the PCM must be significantly less than the output that the heating system provides, making compounds with lower melting points more advantageous.

Organic PCMs are more expensive than inorganic salts but can be custom made to provide transition temperatures specific for a given application. Compared to inorganic salts, organic PCMs are nontoxic, noncorrosive, nonhygroscopic, more stable during repeated phase change cycles, and compatible with most building materials. PCMs have higher latent heat per unit weight, melt without leaving solid phase sediments, and, most important, show negligible supercooling tendencies. Although their potential flammability is a disadvantage, organic PCMs are used to absorb solar heat for air heating in buildings. Eutectic PCMs behave as salt hydrates but can be custom made to meet specific melting points. Costs are considerably higher than those of organic or inorganic PCMs.

PCM modules consist of a series of PCM-filled plastic pipes embedded in insulation material in order

TABLE IX

Transition Temperature and Latent Heat Energy of Common Groups of Phase Change Materials (PCMs)

Group	PCM type	Transition temperature (°C)	Latent heat (Wh/kg)	Thermal conductivity, liquid/solid (W/mk)
Inorganic salt hydrates	Calcium chloride	27–30	47–53	0.54 ^a /1.09 ^b
	Sodium sulfate (Glauber's salt)	32	70	—
	Zinc nitrate	36	41	0.46 ^c /—
	Magnesium nitrate	89	45	0.49 ^d /0.61 ^e
Organics	Polyglycol E400	8	28	0.19 ^f /—
	Polyglycol E600	22	35	0.19 ^f /—
	Octadecane	28	68	—
	Ecosane	37	69	—
	Paraffin 116 (paraffin wax)	48	58	—
	Paraffin 6403 (paraffin wax)	62–64	48–53	0.17 ^g /0.35 ^b
Fatty acids	Palmitic acid	63	52	0.16 ⁱ /—
	Capric acid	32	42	0.15 ^j /—
Organic/inorganic mixes (eutectics)	Mystiric acid (mainly inorg.)	54	52	—
Aromatics	Naphtalene	80		0.13 ^k /0.34 ^l

^a 38.7°C.^b 23°C.^c 39.9°C.^d 95°C.^e 37°C.^f 38.6°C.^g 63.5°C.^h 33.6°C.ⁱ 68.4°C.^j 38.5°C.^k 83.8°C.^l 49.9°C.

to exchange heat with heating systems. In combination with thermal stratification tanks, flat PCM modules with different melting temperatures enhance efficiency because heat conduction at surfaces is eliminated and the stability of the stratification is improved. This method can be implemented with larger centralized thermal storage tanks.

3.3 Other Thermal Storage Systems

A different thermal storage system is based on the thermophysicochemical properties of assorted chemical compounds. For example, cooling is achieved in chemical heating/cooling pipe systems by the endothermic heat of combining solid- or liquid-liquid mixable materials. In solid-liquid pairs, urea-water and ammonium nitrate-water generate endothermic heat of 8.2 kW/kg when the materials are mixed. Similarly, isobutanol-acetonitrile and

1-butanol-acetonitrile produce 0.8 kW/kg. For regeneration, waste heat of 80°C can be used to redistill the mixtures. Evaluation of the most suitable mixing pairs in terms of cold/heat capacity must be implemented for each specific application.

Innovative latent heat storage systems have been developed for ventilation systems using molecular adsorption and desorption on chemical surfaces. With its extensive surface areas, zeolites are ideal materials that adsorb water vapor, effectively releasing a large amount of adsorption heat. This discharging process can heat inlet air from 25 to 100°C by streaming at 6000 m³/h through 7 tons of zeolites and provides a heating power of 95 kW for 14 h per day. During low heat demand, the charging process takes in airstreams of 130°C in order to desorb water molecules from the zeolite surfaces, releasing a humidified airstream at 40°C. The storage capacity is approximately 150 kWh/m³ of zeolite material.

4. CHEMICAL STORAGE SYSTEMS

The energy density achieved with chemical storage is approximately a factor 100 higher than that of any other storage technology discussed previously (Table II). This decisive advantage gives chemical storage systems unsurpassed competitiveness. Primary energy is predominantly stored chemically in large crude oil and natural gas storage tanks or as stockpiles or in silos of specific energy carriers, such as coal or biomass. Secondary or final chemical storage involves refined petroleum products kept in tanks of fuel stations, industries, vehicles, and households. Energy products from coal and biomass, such as briquettes, granulated coal, wood chips, or biomass pellets, are kept dry in sheds, cellars, or silos.

4.1 Coal

Coal is a combustible mineral formed over millions of years through the partial decomposition of plant material subject to increased pressure and temperature under the exclusion of air. Primary coal exists in different grades called anthracite, bituminous, and lignite, with decreasing heating values, respectively. Secondary coal is produced from primary coal into metallurgical coke, anthracite, bituminous, and lignite briquettes. World coal reserves are approximately 9.1×10^{11} tons, with annual consumption of 4.5×10^9 tons.

In 1999, coal production was approximately 5 billion tons worldwide (Table X). Although production stagnated at this level, coal mining

productivity has steadily increased during the past decade.

The key environmental drawback of coal mining is the continuous release of CO₂ due to the glow burn of coal floss. Glow burn is impossible to control because oxygen enters the vast underground coal reservoirs during mining activities, leading to self-ignition. Coal combustion releases particulates and the chemical compounds SO₂ and NO_x, which cause environmental acidification. Pollution abatement equipment enables modern plants to control these emissions. Together with other essential equipment for coal receiving, raw coal storage, coal preparation, coal transportation, and coal storage, this results in significantly higher capital investments.

Coal stocks are kept at electric utilities or at coke and power plants and other industrial facilities. At coal-handling facilities, crushers and granulators use coal delivered from different coal pits to produce coal fragments of approximately 50 mm. The raw coal for power plants or boilers is stockpiled or stored on site in coal silos. The coal is then crushed into powder in pulverizers to feed the plant's burners. Coal storage facilities consisting of coal bunkers, silos, or coal strait or stockpile slot systems allow coal from the different pits to be blended and provide short-term storage to accommodate peak shipping. Customized modeling solutions to deal with high fugitive dusting potentials, storm run-off collection, and settling ponds at ground storage facilities and railroads are necessary. Environmental quality control is maintained using air quality modeling and monitoring, including particulate source inventories, weather observation, and forecasting. In addition,

TABLE X

World Coal Production, 1999^a

Region/country	Thousand short tons			
	Anthracite	Bituminous	Lignite	Metallurgical coke ^b
North America	4768	1,086,789	100,070	26,185
Central and South America	28	49,457	0	10,694
Western Europe	10,308	104,661	380,612	43,499
Eastern Europe and Russia	35,576	475,221	235,399	66,454
Middle East	0	1389	0	335
Africa	2881	324,294	0	5354
Asia and Oceania	298,961	1,757,120	185,329	214,601
World total	352,522	3,798,932	901,410	367,122

^a Source. Energy Information Administration (<http://eia.doe.gov>).

the clean coal technology initiative has led to pollution control and power-generating processes using low sulfur subbituminous coal that reduces air emissions from coal-burning plants, including greenhouse gases. Newer technologies have the potential to utilize high-sulfur “dirty” coal and still achieve comparable air emissions. Enhancing power plant efficiency and capturing carbon gases further reduce CO₂ emissions.

4.2 Crude Oil

World oil reserves are estimated at 1.6×10^{14} liters (1×10^{12} barrels), with potentially comparable amounts of undiscovered oil resources. World consumption is 1.2×10^{10} liters per day. Current oil production is predicted to increase by at least 1 million barrels every year over the short term, with oil reserves remaining economically exploitable for the next 25 years. Prices are expected to increase until 2040 due to the use of costlier new exploitation technologies.

Having adequate storage capacities helps at least to balance short-term price fluctuations. In the long term, however, profitable returns on the hefty investments of oil firms may only be attainable if oil prices steadily rise and additional capital is made available at the right time. Efforts to develop a valid pricing system of as scarce a resource as crude oil seem unattainable due to many uncertainties, including unstable political scenarios. OPEC's current levels of oil production and other factors are major uncertainties impacting price stability. Storage alone provides no answer for these long-term uncertainties.

In the short term, simple market dynamics can utilize primary storage of crude oil to replenish low fuel stocks during mild winters when demand is low for heating oil, propane, and natural gas. In addition to the importance of primary storage for the functioning of the oil market, efficient secondary storage of petroleum products such as gasoline, diesel, jet fuel, other (mostly cracked) oil distillates used for heating, and other residual oil stocks (light and heavy oil) is crucial for corresponding energy services (Table XI).

Many countries regulate oil imports with the use of strategic petroleum reserves (SPRs), with compulsory storage ratios stipulated between importing companies and the government. Large consumer companies can join the contracted storage system under the obligation not to empty their stock below a

given ratio. SPRs play a key role in stabilizing price fluctuations in the international energy markets.

Depending on the annual demand and capital cost of storage and the aging of oil, different amounts of stock are kept by the consumer companies. Above-ground and underground storage tanks are used for secondary storage of high- or low-sulfur distillates. The total fuel distillate ending stock in the United States was approximately 140,000 Mbbl in 2001, with approximately 40% high-sulfur and 60% low-sulfur distillates.

TABLE XI

Example of Crude Oil and Petroleum Storage: Product Stocks^a from the United States in Mbbl (1×10^6 Barrels)^b

Year/product	December 2001 (Mbbl) ^c
Crude oil ^d	312,000
Total motor gasoline	209,000
Reformulated	44,900
Oxygenated	400
Other finished	116,200
Blending components	48,400
Jet fuel	42,000
Distillate fuel oil ^e	144,500
≤0.05% sulfur and under (low)	82,300
>0.05% sulfur (high)	62,200
Residual fuel oil	41,000
Unfinished oils	87,700
Other oils ^f	199,100
Total (excluding SPR)	1,036,100
Crude oil in SPR ^g	550,200
Total (including SPR)	1,586,300

^aProduct stocks include domestic and customs-cleared foreign stocks held at, or in transit to, refineries and bulk terminals and stocks in pipelines. Stocks held at natural gas processing plants are included in “other oils” and in totals. All stock levels are as of the end of the period.

^bFrom the Weekly Petroleum Status Report, Energy Information Administration (<http://eia.doe.gov>).

^cData may not add up to total due to independent rounding.

^dCrude oil stocks include domestic and customs-cleared foreign crude oil stocks held at refineries, in pipelines, in lease tanks, and in transit to refineries. These do not include those held in the Strategic Petroleum Reserve (SPR).

^eDistillate fuel oil stocks located in the Northern Heating Oil Reserve are not included.

^fIncluded are stocks of all other oils such as aviation gasoline, kerosene, natural gas liquids, and LRG's, other hydrocarbons and oxygenates, aviation gasoline blending components, naphtha and other oils for petrochemical feedstock use, special naphthas, lube oils, waxes, coke, asphalt, road oil, and miscellaneous oils.

^gCrude oil stocks in the SPR include non-U.S. stocks held under foreign or commercial storage agreements.

4.2.1 Oil Tanks and the Environment

Oil storage tanks, some with capacities of more than 390,000 barrels (more than 16 million gallons), store crude oil, intermediate stocks (partially refined), finished products, and chemicals. To avert soil and water pollution, containment areas must be designed to hold their capacities in the event of a tank rupture. At refineries, containment systems also divert rainwater runoff to the refinery's effluent treating facilities to ensure that no traces of chemicals or hydrocarbons enter local waters. Processing crude oil can lead to methane venting, a potent greenhouse gas. To reduce methane emissions, liquid storage facilities are consolidated and centralized in the field. From an environmental perspective, all storage tanks should be impermeable to volatile organic compounds (VOCs) and should be equipped with pressure-vacuum valves combinable with a vapor recovery system. Emissions of gas built up in crude oil or produced in water storage tanks are controlled during handling by flare combusting of the VOCs or by tank battery vapor recovery systems. In order to meet air pollution standards, petroleum products are refined to sulfur concentrations as low as 5 parts per million (ppm). The gasoline additive methyl-tertiary-butyl ether must be eliminated gradually due to its persistent properties if leaked into water systems. Tens of billions of dollars are required for replacement and remediation of contaminated groundwater and soil, due to leaking petroleum hydrocarbon tanks in the United States during the early 1990s from secondary and final storage (mostly heating oil and gasoline tanks).

4.3 Natural Gas

World reserves of natural gas are estimated at $1.4 \times 10^{14} \text{ m}^3$, with gross production of gas at $2.4 \times 10^{12} \text{ m}^3$ per year. Natural gas hydrates from deposits in ocean floors contain vast potential, which currently cannot be exploited economically.

Seasonal fluctuation in the consumption of natural gas varies as much as 50% with changing weather conditions and heating and electricity demand. In the short term, more extensive variations can be observed. The volume of peak demand is unpredictable and the exact timing impossible without integrated storage systems. Because natural gas production does not correspond to these fluctuations, surplus gas is injected during low demand into numerous, mostly underground storage facilities for use during high-consumption periods. Underground storage systems are located in porous rocks of depleted fields or aquifers or in

artificial underground caverns, such as abandoned mines or salt caverns. Well-functioning storage systems eliminate the need for expensive, additional pipeline transmission capacity that would be necessary to supply peak demand. In addition, they balance summer production levels with possible supply shortages in the winter.

A gas pipeline network connects supply and demand sides with corresponding storage sites. Storage facilities are located near consumption areas to reduce the occurrence of bottlenecks within gas transmission grids due to gaps between transportation capacities and withdrawals. Natural gas is stored in two basic ways: It is either compressed in tanks as LNG or stored in large underground storage facilities (Table XII). LNG is kept as a chilled and pressurized liquid in double-walled cryogenic vessels or tanks. For transportation by ships or trucks, LNG requires 600 times less storage space than its gaseous form, and its ignition potential is reduced. Natural gas kept in tanks as compressed natural gas is used as vehicle propellant in combustion engines. Volumetric capacity depends on the head and shell thickness of the storage vessel.

Increasing demand, changing demographics, and the deterioration of storage and pipeline systems pose significant technological challenges for the future. Natural gas usage is projected to increase at a rate of 2% per year for the next two decades due to increased substitution of petroleum, diesel, or oil. A significant shift in consumption patterns is expected as more natural gas is used for heating, cooking, and mobility along with the decentralized generation of electricity.

Although natural gas vehicles (NGVs) have penetrated niche commercial markets for buses, delivery vans, and trucks, their success is severely constrained by the cost, weight, and size of existing natural gas storage cylinders. To counter this, a state-of-the-art, safety-certified natural gas storage unit called the integrated storage system (ISS) was developed. The ISS utilizes lightweight, high-density

TABLE XII

Example of Natural Gas Storage^a

Traditional storage	Salt caverns	Total
5927	145	6072

^aSource. Primary storage in the United States for 2000 in volumes of billion cubic feet at 1 bar pressure and 15°C. Energy Information Administration, *Natural Gas Monthly* 2001 March (http://www.eia.doe.gov/oil_gas/natural_gas/nat_frame.html).

polyethylene in the construction of small-diameter cylinders (pressure cells), which are subsequently encapsulated within a high-strength fiberglass shell containing impact-absorbing foam. The total energy storage capacity of 45 liters of gasoline equivalent at a service pressure of 250 bar enables a NGV to travel more than 480 km before refueling. This is comparable to the distance traveled by vehicles powered by gasoline. ISS technology reduced the cost of NGV storage tanks by approximately 50%. Its weight savings is up to 70% compared to steel and aluminum cylinders, and ISS does not interfere with the car's crumple zone. For commercial viability, the cost of storage must be reduced.

4.4 Hydrogen

Hydrogen is mainly used as chemical feedstock in the petrochemical industry and in the food, electronics, and metallurgical processing industries. However, hydrogen is believed to have great potential as a major energy carrier for clean energy systems of the future.

Hydrogen can be used in transportation, construction, utilities, and industry as a chemical storage compound. It can be used to mitigate CO₂ emissions by upgrading carbon-rich fuels to less carbon-intensive fuels. Renewable energy resources may be used to extract hydrogen from water, which is then combusted by fuel cells, for example, reforming water and balancing the intermittent nature of renewables. Ideally, combusting hydrogen generates no atmospheric emissions other than water.

For vehicles, hydrogen would be ideal as a synthetic fuel because it is light and highly abundant. The hydrogen fuel could be stored in carbon composite tanks at pressures of up to 250 bar, providing fuel for more than 200 km. Provenzano *et al.* estimated that if a hydrogen combustion engine operates in lean burn mode, the vehicle produces no CO and only low NO_x emissions (10–100 ppm). Research is directed at developing hybrid zero-emission vehicles that use hydrogen for lean-burning, spark-ignited engines combined with flywheels, supercapacitors, or batteries as energy storage devices. Mobility based on hydrogen could become competitive with continued technological development and cost reductions. Regenerative energy storage for ultra-high-flying solar aircraft is another use for a hydrogen support system. The system was designed by NASA to store energy by converting the craft's excess solar power into hydrogen and oxygen during the day and then to use a fuel cell to "regenerate" electricity at night.

The hydrogen storage necessary to continuously power automobiles or fuel cells is a major safety problem. Hydrogen is stored as a compressed gas in vehicles or as a cryogenic liquid in physical storage systems for transport. The solid-state storage system is another, safer storage technology that could potentially store more hydrogen per unit volume. Solid-state systems bind hydrogen to a solid material such as carbon. Critical parameters for solid-state storage are hydrogen uptake capacities, controlled hydrogen release, and rapid hydrogen refilling. Potential advantages of solid-state storage are its applicability at higher temperatures and near ambient pressures. According to Schlapbach and Züttel, lightweight metal alloy hydrides are the most promising materials for hydrogen storage.

4.5 Biomass

Biomass fuels (biofuels) refer to chemical energy stored within plants by photosynthesis. Theoretically, 6.7% of the solar energy input on land could be converted to chemical energy by photosynthesis, although on average only 0.3% is stored as carbon compounds in land plants. Khesghi *et al.* calculated net primary productivity to be 60 GtC per year on land and 50 GtC per year at sea. With different heating values, biomass varies in both its energy and its carbon content.

Primary biomass storage includes any organic matter available on a renewable basis. It comprises biomass stocks and an assorted range of waste and residue materials. The basis for secondary storage before distribution and final or end-use storage is a continuous biomass resource supply chain to produce solid and liquid biofuels or synthetic gases (syngas).

The chemical energy released as heat by combustion is mostly used for room heating or industrial processes. It can also be converted to electricity. Gaseous or liquid fuels such as methane, hydrogen, and alcohols or bio-oils can be produced from biomass instead of direct heat production. Liquid biofuels and biogas are applicable in IC engines or fuel cells as alternatives to petroleum-based fuels. The conversion of biomass to electricity is conceptually similar to the conversion of coal to electricity using either direct combustion or integrated gasification combined cycle technology.

The carbohydrates of biomass can be converted by fermentation to ethanol or by extraction and refining to plant oils. For example, corn can be converted to a fermentation feedstock, where microorganisms

including yeast and other bacteria produce ethanol, which can be refined and used for blended gasoline, an alternative transportation fuel. Another example is the reaction of plant oils (fatty acids) with methanol in the presence of a catalyst, producing biodiesel. Biodiesel is used as an alternative fuel especially in agriculture either in neat form or blended with petroleum diesel for use in diesel engines.

One major issue with regard to secondary or final storage of biofuel is the oxidative stability of the compounds. Stability test methods allow a customer to determine whether the fuel has remained stable in storage over extended periods of time. Biodiesel stored in standard diesel storage tanks interacts with the accumulated tank sediments and sludge from ordinary diesel and reduces the solvency problem of ordinary diesel. For solid fuel storage, security measures are vital for preventing (self)-ignition in the presence of heat and oxygen.

Biogas is the most promising biofuel. It is more directly accessible, has good storage stability, has better overall energetics, and can be applied more flexibly. It can also be produced by fermentation processes, pyrolysis, or gasification techniques from diverse biomass raw materials. Biogas from 1 kg of air-dried, gasified biomass stores approximately 8000 kJ of energy, which can deliver approximately 0.8 kWh of electric output. Biogas production from biomass pyrolysis takes place at high temperatures, under exclusion of oxygen in the presence of catalysts. Biogas storage is no different from the storage of natural gas. Since biogas contains hydrogen sulfide, tank materials must be H₂S resistant. The purity of biogas compared to natural gas is a controversial issue in debates on the economic feasibility of large biogas production plants, which would require hefty cost reductions and efficient conversion technologies.

The central problems in producing bioliquids are its economic cost and the negative energetics of production. Besides growing and harvesting the biomass stock, significant amounts of energy are consumed in the conversion of biomass to biofuels. Approximately 6% of the (lower) biomass heating value is used per unit production. Developments to improve the overall efficiency of biomass energy systems will be the key to making this a feasible alternative. Another distinct characteristic of biomass energy sources is the large terrain requirements for generating biomass stock. Terrestrial plants, however, are not the only option as renewable energy stock. Marine algae, including kelp, are potential

renewable energy resources but are currently too expensive to utilize.

5. CONCLUSION

Although an efficiency loss is unavoidable for all storage systems, the capacity and services of energy systems, including load distribution and system reliability, must continually be optimized. Load-distribution strategies and energy end-use efficiency increase customer benefits, whereas energy storage and control devices allow consumers to time energy purchases and shift from high- to low-rate periods. Energy storage also increases system reliability and will be the key to balancing the intermittency of new renewable energy sources in the future.

Efficient electricity storage in larger quantities is only indirectly possible by using established storage techniques, such as pumped storage hydropower plants pumping water back to the reservoir at times of low electricity demand. For such large systems, the foremost limiting factor for future growth is the availability of suitable sites with low environmental impact. Similarly, the development of large-scale CAES will depend on the development of systems with man-made storage tanks.

During the past 20 years, smaller scale electricity storage technologies have advanced greatly. New energy storage systems involving flywheels are very promising, especially in combination with other storage technologies to support uninterruptible electricity supply systems, thus improving reliability and quality. Batteries remain significant for smaller appliances and applications in which grid power independence is necessary. Supercapacitors have been developed as an additional electricity storage technology to maintain grid stability and satisfy rapid power needs.

Thermal storage technologies are especially useful in combination with solar energy, CHP, and waste heat distribution to provide hot water and heat for housing. Although large thermal storage systems such as high-temperature UTES and ATES are more efficient, small-scale applications intelligently integrated into building technology have immense potential.

Chemical storage remains the definitive storage method, with its high energy density. The use of coal and oil, however, poses great problems for sustainable energy due to CO₂ emissions. The transition to natural gas, hydrogen, and biomass as prime chemical energy storage is envisaged for the future. Cost

reductions in the various chemical storage techniques will be crucial for applications in mobility, distributed electricity generation, and heat production.

Energy storage should be lean, safe, and reliable to contribute to the functioning of energy supply markets and energy systems. Some redundant capacity of energy storage is important for all sectors of the socioeconomy to cope with disruptions caused by various technical, economical, or political events.

Acknowledgments

I thank D. Spreng (CEPE) for his encouragement and advice with regard to this article. I am indebted to J. Appleby (Texas A&M University) for invaluable comments and corrections. I also thank M. Berg (Factor AG, Switzerland) for his valuable comments on the fossil fuel sections. Thanks also to D. Goldblatt for reading the manuscript. Finally, I thank Anne Semadeni-George for her extraordinary help in reducing the text to an acceptable length.

SEE ALSO THE FOLLOWING ARTICLES

Coal Storage and Transportation • Hydrogen Storage and Transportation • Natural Gas Transportation and Storage • Strategic Petroleum Reserves • Thermal Energy Storage

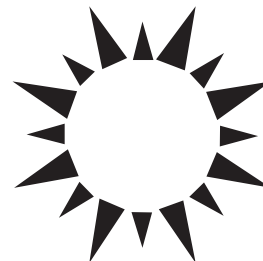
Further Reading

- Aceves, S. M., and Smith, J. R. (1996). "Hybrid and Conventional Hydrogen Engine Vehicles That Meet EZEVE Emissions." Lawrence Livermore National Laboratory/U.S. Department of Energy, Livermore, CA/Washington, DC.
- Banks, F. E. (2000). "Energy Economics: A Modern Introduction." Kluwer, Boston.
- Blaisdell, R. A., and Smallwood, M. E. (1993). "Evaluation of the Total Hydrocarbon Standard for Cleanup of Petroleum Contaminated Sites." Wright-Patterson Air Force Institute of Technology, OH, USA.
- Chartier, P., Beenackers, A. A. C. M., and Grassi, G. (1995). "Biomass for Energy, Environment, Agriculture and Industry." Pergamon, Oxford, UK.
- Collinson, A. (2001). "Daces 2050: Database Clean Energy Supply 2050." Netherlands Ministry of Housing, Spatial Planning and the Environment/Utrecht Centre for Energy Research/Department for Science, Technology and Society, Utrecht University/Netherlands Energy Research Foundation (ECN)/Ecofys, Energy and Environment/National Institute of Public Health and the Environment; Utrecht, The Netherlands.
- Dinçer, I., and Rosen, M. A. (2002). "Thermal Energy Storage: Systems and Applications." Wiley, Chichester, UK.
- Dresselhaus, M. S., and Thomas, I. L. (2001). Alternative energy technologies. *Nature* **414**, 332–337.
- EESAT. (1998). Proceedings of the International Conference of Electrical Energy Storage Systems: Applications and Technologies. In "International Conference on Electrical Energy Storage Systems," Chester, UK.
- Ewald, R. (1998). Requirements for advanced mobile storage systems. *Int. J. Hydrogen Energy* **23**(9), 803–814.
- Howell, S. A., and Weber, J. A. (1996). "U.S. Biodiesel Overview." National Biodiesel Board, Jefferson City, MO.
- International Energy Agency (1997). "Implementing Agreement on Energy Conservation through Energy Storage." International Energy Agency, Paris.
- International Energy Agency (2000). "Agreement on the Production and Utilization of Hydrogen." International Energy Association/C. C. Elam National Renewable Energy Laboratory, Paris/Golden, CO.
- International Energy Association (2001). "Implementing Agreement for Hydropower Technologies and Programmes—Annex III: Hydropower and the Environment: Present Context and Guidelines for Future Action." International Energy Agency, Paris.
- Khesghi, H. S., Prince, R. C., and Marland, G. (2000). The potential of biomass fuels in the context of global climate change: Focus on transportation fuels. *Annu. Rev. Energy Environ.* **25**, 199–244.
- Larbalestier, D. (2001). High-Tc super-conducting materials for electric power applications. *Nature* **414**, 368–377.
- Linden, D., and Reddy, T. B. (2002). "Handbook of Batteries." McGraw-Hill, New York.
- Lorén, A., Hallbeck, L., Pedersen, K., and Abrahamsson, K. (2001). Determination and distribution of diesel components in igneous rock surrounding underground diesel storage facilities in Sweden. *Environ. Sci. Technol.* **35**(2), 374–378.
- Masuda, M. (1988). Lecture on new technology for energy conversion and its storage. Storage of energy, part 2: Storage of electric energy by super-conductivity. *Nenryo-Kyokai-Shi J. Fuel Soc. Japan* **67**(9), 795–801.
- Maxwell, R. M., and Kastenberger, W. E. (1999). A model for assessing and managing the risks of environmental lead emissions. *Stochastic Environ. Res. Risk Assessment* **13**, 231–250.
- Owen, A. D. (2001). Conference Proceedings 2001: The Economics of Renewable Energy Technologies. In "24th IAEE International Conference. Energy Odyssey?," held in Houston, TX. Omnipress, Madison, WI.
- Padró Grégoire, C. E. (2000). "Advances in Hydrogen Energy." Kluwer /Plenum, New York.
- Provenzano, J., Scott, P. B., and Zweig, R. (1997). Solar hydrogen for urban trucks. In "8th Annual U.S. Hydrogen Meeting Proceedings," pp. 209–221. National Hydrogen Association, Washington, DC.
- Ristinen, R. A., and Kraushaar, J. J. (1999). "Energy and the Environment." Wiley, New York.
- Schlapbach, L., and Züttel, A. (2001). Hydrogen-storage materials for mobile applications. *Nature* **414**, 353–358.
- Turner, W. C. (2001). "Energy Management Handbook." Fairmont, Lilburn, GA.
- Witt, E. C., Mitchell, R. L., Holly, T. P., and Symko-Davies, M. (2001). Terrestrial photovoltaic technologies update. *Renewable Energy* **23**(3/4), 349–353.
- Yeager, K., Gehl, S., Barker, B., and Knight, R. (1998). Roadmapping the technological future of electricity. *Electricity J.* **11**(10), 17–31.

Strategic Petroleum Reserves

DAVID L. WEIMER

University of Wisconsin
Madison, Wisconsin, United States



-
1. Importance of Strategic Petroleum Reserves
 2. Historical Overview
 3. Rationale for Government Role
 4. Economic Impacts of Stock Draw Downs
 5. Costs of Strategic Petroleum Reserves
 6. Time and Risk
 7. Programmatic Approaches
 8. Conclusions

Glossary

consumer surplus The monetary value individuals place on consumption in excess of the amount of money they have to pay to obtain it.

draw down The release of petroleum from strategic stocks.

in situ storage Holding crude oil stocks in the ground by capping wells.

marginal excess tax burden The additional cost to the economy of another dollar of taxation.

market failure A situation in which the self-interested actions of market participants does not lead to an economically efficient allocation of resources.

mmb/d Millions of barrels per day.

oil price shocks Steep increases in the price of crude oil in the world market resulting from the sudden removal of supply from the market.

price elasticity of demand The percentage change in the quantity of a good demanded in a market resulting from a 1% change in its price.

price elasticity of supply The percentage change in the quantity of a good supplied to a market resulting from a 1% change in its price.

primary stocks Petroleum held by crude oil producers and refiners. Secondary stocks are those held by wholesalers and retailers; tertiary stocks are those held by consumers.

producer surplus Payments to producers of a good in excess of the minimum amount required to produce it.

risk aversion A preference for a certain payoff over a lottery with an expected outcome equal to the certain payoff.

salt domes Caverns in underground salt formations that can be created through the injection of fresh water and

the removal of the salt in solution and used to store petroleum.

speculative stocks Petroleum accumulated in expectation of profitable sales at higher future prices.

strategic stocks Petroleum stocks held either to facilitate draw downs to reduce the economic impacts of supply disruptions or to deter purposeful reductions in supply for political ends.

supply disruption The sudden removal of substantial quantities of crude oil from the world market. Major disruptions occurred in conjunction with the Yom Kipper War (October 1973 to March 1974), the Iranian Revolution (November 1978 to April 1979), the Outbreak of the Iran-Iraq War (October 1980 to January 1981), and the Gulf Crisis (August 1990 to January 1991). The largest of these in magnitude was the Iranian Revolution, which removed 5.6 mmb/d for approximately 6 months.

surge capacity The incremental supply that can be added to the market over relatively short periods of up to 6 months.

working (operational) stocks Petroleum held to facilitate normal industry operations, including stocks in pipelines, tankers, and tank bottoms, stocks to smooth tanker deliveries, and stocks to accommodate seasonal demands.

Strategic petroleum reserves, also referred to as strategic petroleum stocks, are quantities of crude oil or petroleum products held either to facilitate draw downs to reduce the economic impacts of crude oil supply disruptions or to deter purposeful reductions in crude oil supply for political ends. They reduce the vulnerability of the world economy, especially the economies of petroleum importers, to the adverse consequences of sudden and steep price increases resulting from the removal of substantial quantities of crude oil supply from the world market. They make purposeful reductions in crude oil exports for political purposes less attractive by increasing the quantities that must be withheld to inflict any threatened level of harm on petroleum importers. In these ways, strategic petroleum reserves reduce the

risks to the world economy associated with reliance on geographically concentrated crude oil supplies.

1. IMPORTANCE OF STRATEGIC PETROLEUM RESERVES

A healthy world economy requires a steady flow of crude oil. In the short run, the relatively small absolute value of the price elasticity of demand for crude oil means that even modest reductions in supply can result in dramatic increases in price. Such price shocks result in direct losses to the national economies of net importing countries and produce adverse macroeconomic effects for most of the rest of the world. Indeed, economists attribute much of the poor economic performance of the world economy during the 1970s to the oil price shocks occurring in connection with the Yom Kippur War, which quadrupled world prices, and the Iranian Revolution, which doubled them. Continued dependence on petroleum as a major energy source will keep the world economy vulnerable to crude oil supply disruptions into the foreseeable future.

The risk of oil supply disruptions will remain high because of the concentration of crude oil production and reserves in the volatile Persian Gulf region. In 2002, the world consumed on average about 77 mmb/d of crude oil. Approximately 29% of this supply originated in the Persian Gulf. With more than half of the world's proved oil reserves, the Persian Gulf is likely to account for an even larger fraction of supply in the coming years. The disorder in Iraq and longer term concerns about the stability of the Saudi Arabian and Iranian regimes make it easy to imagine scenarios in which the world would lose some or all of the oil exported from the Persian Gulf. For example, a revolution in Saudi Arabia that damaged export facilities would deny the world oil market approximately 10 mmb/d and, without offsetting increments of supply, send the world price of oil to over \$50/b. Less severe disruptions could originate in other parts of the world, such as the countries of the former Soviet Union. The primary purpose of strategic petroleum reserves is to provide a source for offsetting increments of supply in such situations.

2. HISTORICAL OVERVIEW

The holding of commodity stocks is one of the oldest functions of government—recall the biblical story of

Joseph advising the Pharaoh to put away the surplus of the coming seven good years to be prepared for the seven lean years to follow. In the early twentieth century, oil came to top the list of critical resources as navies converted from coal to oil and economies became motorized. By the end of the century, civilian use came to overshadow military needs so that oil stocks came to be viewed as more important for economic than military security.

2.1 Naval Petroleum Reserve

The conversion of the U.S. Navy to oil-burning ships made access to petroleum an immediate national security concern. Even though the United States was then a net exporter of crude oil, beginning in 1912 the United States set aside oil-bearing lands in California, Wyoming, and Alaska as a naval petroleum reserve. The reserve was to ensure that crude oil would be available for use by the navy in the event of an extended military conflict.

The naval petroleum reserve can be thought of as *in situ* storage of strategic stocks. Rather than removing oil from the ground and placing it in storage, the oil is stored in place. Although *in situ* storage seems like a commonsense approach to creating strategic reserves, the historical experience with the naval petroleum reserve reveals several problems with storing strategic stocks in the ground.

First, unless the land area of the reserve is sufficiently large to preclude other land owners from tapping the same reservoir of oil, production must be undertaken to protect the reserve from commercial encroachment. This has been the case at one of the naval petroleum reserve sites since the 1920s, as well as intermittently at two of the other sites, including the infamous Teapot Dome. The problem was particularly severe in the U.S. context where the rule of capture gives ownership of oil and gas to anyone who brings it to the surface on their own land. Yet preserving the reserve would likely be a problem whenever the entire reservoir is not dedicated as strategic stocks.

Second, *in situ* reserves are often not located near transportation facilities with adequate capacity to handle surge production. In the case of the Alaskan Naval Petroleum Reserve, for example, a second Alaskan oil pipeline would have to be built and then left to stand idle if the oil from the reserve were to be available as additional supply to the market. Providing surge capacity places a high demand on transportation facilities.

Third, it is generally costly and difficult to keep shut-in wells in a state of readiness that allows them to achieve full levels of production rapidly. After the Six Day War in 1967, the secretary of the navy ordered one of the naval petroleum reserve sites to be prepared to produce at its maximum expected rate within 60 days. The readiness requirement was never met, and, when commercial production was begun in 1976, the maximum rates of production anticipated turned out not to be achievable.

In situ storage like that provided by the naval petroleum reserve has an even more fundamental disadvantage: the large amount of proved reserves that must be set aside to achieve substantial surge capacity. It is rarely possible to extract more than one-eighth of a reservoir within 1 year. Imagine that the United States wanted to have the capability for 3 mmb/d of surge capacity for 1 year. Reservoirs of approximately 8 billion barrels of proved reserves would have to be set aside, more than one-third of all U.S. proved reserves. Once the purpose of strategic oil stocks shifted from supporting military operations in extended conflicts to providing surge capacity for avoiding the economic losses from oil price shocks, *in situ* stocks were no longer appropriate as strategic stocks.

2.2 Stockpiling as a Response to Import Dependency

Several policies limited U.S. dependency on imported oil through the early 1960s. One was an oil import quota that attempted to limit imports directly. Initiated in 1957, it remained in effect until rising world prices and controlled domestic prices made it politically unworkable in 1973. The other was the pro-rationing of production in unitized oil fields by state regulatory agencies in Texas and Oklahoma. For example, the Texas Railroad Commission attempted to stabilize crude oil prices by raising and lowering the pro-rationing rate in response to changes in demand. By 1972, however, prorating reached 100%, removing the capacity of the state agencies to stabilize price.

The first proposals for holding strategic reserves came from economists who distinguished between import dependence and vulnerability. They argued that it was not import dependence per se that was worrisome, but rather that the imports were subject to disruption from world events. From this perspective, energy security could be achieved by reducing vulnerability even with substantial reliance on

imports. These ideas began to take hold politically, but legislation to create a strategic petroleum reserve introduced in the Senate in April 1973 failed to pass.

2.3 International Energy Program

In October 1973 the Organization of Arab Petroleum Exporting Countries (OAPEC), a subset, including the most significant members, of the Organization of Petroleum Exporting Countries (OPEC), imposed an embargo on oil shipments to the United States and The Netherlands in retaliation for their support of Israel in the Yom Kipper War. At the same time, OPEC cut total production. Although the flexibility of the world oil market kept the United States and The Netherlands from suffering disproportionately from the embargo, the reductions in oil supply dramatically increased oil prices in the world market. In response, 16 of the then 24 members of the Organization for Economic Development and Cooperation (OECD) adopted the International Energy Program to protect its members from targeted embargoes and prepare for future oil supply disruptions. A formal oil sharing agreement was included to counter targeted embargoes, and requirements for national-level programs of demand restraint and the establishment of strategic petroleum reserves were included to reduce the impacts of supply disruptions.

The International Energy Program set a stock requirement for its members based on levels of imports. Specifically, each country is required to hold petroleum stocks, exclusive of military stocks and oil in transit at sea, equivalent to 90 days of net petroleum imports. (Members of the European Economic Community face a similar requirement based on consumption, with partial waivers for Britain and Denmark, its net oil exporters.) Countries may count as much as 90% of working stocks toward the requirement. In 1995, the members of the International Energy Program agreed to coordinate their draw downs of strategic stocks during oil supply disruptions, and to do so before initiating sharing arrangements.

The capability provided by the International Energy Program's stockpiling can be measured in terms of the maximum amount of petroleum that can be delivered to the world market at the onset of an oil supply disruption. Member country governments control public petroleum stocks that would enable them in an all-out coordinated draw down to release perhaps as much as 12 mmb/d of stocks during the first month of an oil supply disruption, followed by approximately 8 mmb/d for the next 2 months. In

subsequent months the volume of the draw down would decline steeply, reaching negligible levels by the eighth month. Some additional releases from commercial stocks could further augment the draw down, especially after prices peak.

3. RATIONALE FOR GOVERNMENT ROLE

Private firms routinely hold petroleum stocks. Transporters and refiners hold primary stocks to maintain the transportation system (tank bottoms and pipeline fill), to avoid production interruptions from delivery delays, and to smooth out seasonal demands. Wholesale and retail firms hold secondary stocks to ensure that they can meet customer demands. End users of petroleum products may hold tertiary stocks to reduce the consequences of delayed deliveries. In general, these are working, or operational, stocks in the sense that they provide the grease needed to keep the distribution system working smoothly. Beyond these routinely held stocks, firms may seek profits by buying speculative stocks at low prices in anticipation of future higher prices. Specifically, the holding of a barrel of speculative stocks will be profitable in an expected value sense if the percentage increase in the expected price less storage costs over the current price is greater than the rate of interest. For example, if the expected price 1 year from now were \$35/b with physical storage costs to be paid at that time of \$2/b, and the current price were \$30/b, it would be profitable in an expected value sense to add to stocks if the interest rate facing the firm were 10% or lower. (The firm could earn \$33 dollars by simply investing the \$30 at the 10% interest rate instead of holding the stocks.) Does private speculation provide adequate strategic stocks from the perspective of society as a whole? In other words, are there market failures that provide a rationale for government stockpiling?

Policy analysts have identified a number of reasons why speculative stocks are likely to be too small to serve adequately as strategic stocks. First, private firms do not gain the full social benefits from draw downs. When firms draw down previously accumulated stocks to profit from higher prices, they produce an external benefit to society—lower prices for petroleum products than would have resulted in the absence of their draw down. This divergence between private and social benefits results in too little speculative stockpiling.

Second, private firms do not see the full social costs from accumulating stocks. When firms accumulate stocks, they do not take account of the social costs of the higher prices that result. Although this divergence could lead to too little speculative stockpiling, the likely timing of additions to stocks may actually add to the social costs of oil supply disruptions. Speculative stockpiling is most attractive to individual firms when they anticipate that prices will rise quickly. This is likely to be the case at the outset of supply disruptions as firms see events leading to supply reductions and attempt to add to stocks. This additional demand may make the price shock associated with the supply disruption steeper and more severe, adding to its social costs. In effect, the divergence between private and social costs is likely to lead to speculative stockpiling that increases rather than decreases social costs.

Third, the model of speculative stockpiling assumes that firms are risk neutral in the sense that they are indifferent between a certain payoff and a lottery with an expected value equal to the certain payoff. So, for example, a firm would be risk neutral if it were indifferent between a zero certain payoff and a lottery that gave a 10% chance of \$9 payoff and a 90% chance of a \$1 loss—although the expected payoff from the lottery is zero, 90% of the time the firm would lose money. It is quite likely that firms would exhibit some degree of risk aversion, which would lead them to choose lotteries only when their expected payoffs are substantially greater than zero. As speculative stocks are like lotteries, risk aversion would likely lead to less speculative stockpiling than would occur if the firms were risk neutral. From the social perspective, however, risk aversion calls for more stockpiling, with the small certain costs of accumulating stocks like a premium for insurance that reduces the large costs that result from oil price shocks.

Fourth, firms may anticipate that government interventions may prevent them from realizing profits from sales of stockpiled oil during price shocks. Throughout the 1970s, U.S. policies controlled the well-head prices of domestically produced crude oil and accumulated stocks. Although these price controls were fully removed in 1981, U.S. firms may anticipate the possibility of reimposition of such controls during a severe oil price shock in the future. This reduces the expected price of speculative stocks by reducing the probability of realizing extremely high prices. Of course, the prospect of the release of public stocks during price shocks also reduces the probability of realizing high prices and thus reduces

speculative stockpiling. At the same time, however, the existence of public stocks may make it less likely that the government will impose price controls and mandatory allocations, making speculative stockpiling more attractive.

Externalities, risk aversion, and fear of government expropriation make a strong case that speculative stockpiling alone will not provide sufficient strategic stocks from the social perspective. Thus, there is a strong rationale for public policy to increase the quantity of strategic petroleum reserves. Determining the appropriate policies for strategic stocks requires an assessment of their social costs and benefits.

4. ECONOMIC IMPACTS OF STOCK DRAW DOWNS

The economic benefits of strategic petroleum reserves accrue when they are drawn down during oil supply disruptions to reduce the magnitude of the resulting oil price shock. The economic impacts can be illustrated by considering hypothetical but plausible price and quantity changes in the world and the U.S. domestic crude oil markets (Fig. 1). The intersection of supply (S) and demand (D) in the world market, which yields an equilibrium price of \$28/b and quantity of 76.8 mmb/d, represents a normal market period. At the world price of \$28/b, the United States consumes 15.0 mmb/d and produces 5.8 mmb/d domestically. The difference between consumption and domestic production, 9.2 mmb/d, is imported. Now imagine that the 10 mmb/d produced by Saudi Arabia is removed from the market. The world supply curve shifts from S to S_d , inducing a price shock that raises price from \$28/b to \$64/b. In the U.S. market, the \$64 price reduces the amount consumed and increases the amount domestically produced.

The net loss to the U.S. economy of the price increase is given by the area of the trapezoid inscribed by the points $cdfe$. It results from a consumer surplus loss equal to the area of trapezoid $bdfa$ less producer surplus gains given by the area of trapezoid $bcea$. The total economic loss from the price shock can be divided into three components. First, consumers suffer a loss from consuming less—area dfh . Second, consumers transfer wealth to foreign oil producers equal to the area $cdhg$. Third, domestic producers spend real resources, equal to area cge , to increase production.

Now imagine that the United States, Germany, and Japan collectively draw down their national

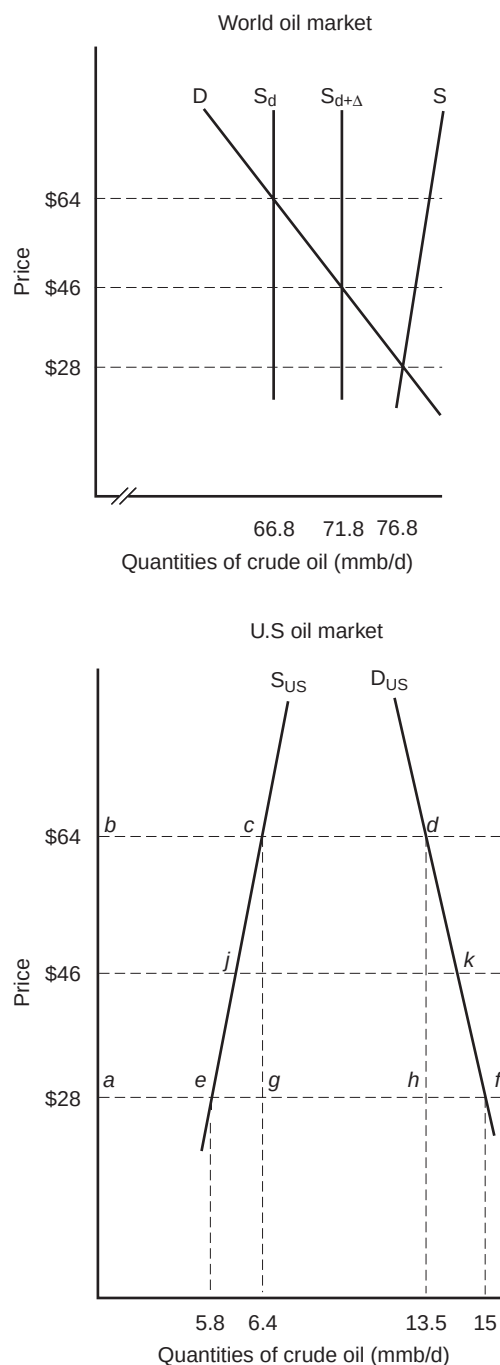


FIGURE 1 The economic impacts of strategic stock draw downs in the world and U.S. crude oil markets.

reserves at a rate of 5 mmb/d. This has the effect of shifting the world supply curve from S_d to $S_{d+\Delta}$. This additional supply in the world market drives price down to \$46/b. The benefits of this draw down in the U.S. market can be seen as the difference between the surplus loss without the draw down (area $cdfe$) and the surplus loss with the draw down (area $jkfe$). The

direct net benefits of the draw down to U.S. consumers is therefore the area of *cdkj*, which equals approximately \$140 million per day, or approximately \$12.6 billion over 90 days.

The illustration makes clear that price is determined in the world market. Thus, draw downs are most appropriately thought of as affecting the world market. Draw downs by any country produce a benefit to petroleum consumers worldwide. Attempts to limit the benefits of draw downs domestically are doomed to failure, as relatively lower domestic prices would divert imports from the country to the higher priced world market, eventually equalizing price. The country drawing down does obtain a direct financial benefit through sales of the drawn down petroleum.

Three other points related to the illustration are relevant to the question of when strategic petroleum reserves should be used. First, efforts by private firms to add to speculative stocks as events lead them to anticipate the higher prices that will result from the supply disruption could be shown in the diagram as increases in demand. In terms of the figure, the speculative stockpiling would shift the demand curve, *D*, to the right, causing price to rise even higher than would be the case with only the loss of supply to the world market.

Second, the economic costs measured in the U.S. market by the net changes in consumer and producer surpluses do not include the macroeconomic costs that result from moving from low to high price equilibria. Interpretation of the surplus changes as economic effects assumes that prices of all other economic inputs and goods freely adjust to the higher price of petroleum. In most economies, however, there are price rigidities, such as labor contracts, that prevent prices from rapidly adjusting. Without free movement of prices, resources will not necessarily be used efficiently. For example, wage rigidities are likely to lead to higher rates of unemployment during oil price shocks. Uncertainty about the composition of post-disruption demand may depress investment. Overall, these additional costs are likely to grow more than proportionally with the size of the price shock, probably approximating the magnitude of the direct costs for a doubling of price.

Third, the illustration is a snapshot showing the initial impacts of the supply disruption. Over time, adjustment by producers, holders of speculative stocks, and consumers will reduce the price of petroleum even before the disrupted supply is returned to the world market. Oil producers will find it profitable to increase production from existing

fields, say by using more expensive recovery methods or reopening previously closed stripper wells, which will tend to rotate the supply schedule, *S*, clockwise around its predisruption equilibrium. Once it appears that price has peaked, those holding speculative stocks will release them, adding to the available supply and removing any shift in demand caused by their accumulation efforts. Petroleum consumers will find ways of reducing petroleum consumption that will tend to rotate demand counterclockwise around the predisruption equilibrium. For example, during the first week of the disruption, consumers can turn down thermostats, make sure that their automobile tires are properly inflated, and take fewer trips; during the first month, they can obtain tune-ups for their automobiles, augment window insulation, and wrap water heaters in insulation; during the first 6 months, they can make more substantial changes in insulation, thermostats, appliances, and vehicles. Consequently, other things equal, the longer the disruption persists, the more market forces, both in terms of incremental supply and reduced demand, will tend to move the price of oil toward the predisruption equilibrium.

Each of these three points argue for large draw downs at the outset of supply disruptions. Immediate draw downs are likely to suppress speculative stockpiling, thus avoiding the additional price increase it causes. The disproportionate increase in macroeconomic costs with the size of the price shock argue strongly for using available strategic stocks to limit the initial magnitude of the price shock. The price mitigating market responses, as well as macroeconomic adjustments, that occur over time suggest that stocks will be relatively more valuable during major supply disruptions if used to support larger initial draw downs than to support smaller but longer draw downs.

5. COSTS OF STRATEGIC PETROLEUM RESERVES

Before strategic petroleum reserves can be drawn down, they must be accumulated and stored. In assessing costs, it is important to distinguish between budgetary costs and social costs. The former are the flows of funds from the national treasury to purchase resources for stockpiling; the latter are the economic value given up by use of the resources for stockpiling. Although budgetary cost is often politically relevant, social costs should be the basis for public policy.

Consider first government-controlled strategic stocks. Their budgetary costs include expenditures to construct (or purchase) storage facilities and maintain them, to connect storage facilities to oil distribution networks, and, most significantly in terms of magnitude, to purchase petroleum. An offsetting revenue is realized when the stocks are sold during drawn down. These budgetary costs may differ from social costs in a number of ways, however.

If petroleum purchases for strategic stocks drive up prices, then they will inflict economic losses on consumers. The economic costs of the price increase can be measured in the same way as price increases resulting from supply disruptions (Fig. 1). Generally, increments to strategic stocks are small relative to world consumption so that they generally will result in negligible, if any, price increases. The major exception occurs at the onset of supply disruptions. (Indeed, stopping ongoing purchases have the same effect as a draw down equal to the magnitude of the purchases.) If purchases are reduced or suspended when petroleum markets are experiencing upward price pressure, then the budgetary and social costs of oil purchases will closely correspond.

The taxes governments use to raise revenues create economic inefficiencies by distorting economic behavior. Economists refer to these inefficiencies as the marginal excess burden of taxation. In assessing the social costs of stockpiling, the marginal excess burden should be applied to revenue flows into and out of the national treasury. For example, if the marginal excess burden for the tax system of a country is 10%, a billion dollars of oil purchases would have a social cost of \$1.1 billion and \$20 billion dollars of revenue from a draw down would yield a social benefit of \$22 billion.

A large discrepancy between budgetary and social costs results when the government bypasses market purchases to obtain oil. For example, the United States receives royalties of between 12.5 and 16.7% of oil obtained from leases on the federally owned outer continental shelf. Usually these royalties are collected as cash equivalents. Plans, however, call for receiving this oil in-kind and using it to fill the Strategic Petroleum Reserve. Although these royalties no longer pass through the budget, they have the same social costs as if they were purchased on the market.

A similarly large discrepancy between budgetary and social costs results when the government mandates that private firms hold strategic stocks. The budgetary costs for purchasing and storing oil

will be zero, but the social costs include the purchase and storage costs borne by the firms.

Storage costs can vary greatly. Storage facilities vary greatly in their cost. Aboveground tanks typically cost between \$15 and \$18 per barrel to build and maintain. In contrast, storage in salt domes costs between \$1.50 and \$5.00 per barrel to build and maintain. Although abandoned salt mines may be available, salt dome capacity is generally created through solution mining, which involves injecting fresh water to dissolve salt that is removed in brine. Temperature differences in deep salt dome facilities provide natural circulation of stored petroleum. The abundance of salt domes on the Gulf Coast near major crude oil pipelines and tanker terminals provided the United States with opportunities for relatively low cost storage and transportation of strategic petroleum reserves.

Maintenance costs for crude oil stocks are generally low. They can be substantially higher for petroleum products, which tend to deteriorate in quality over time and therefore must be rotated out of stocks. The higher maintenance costs if petroleum product stocks has been one factor leading governments to hold mostly crude oil in government owned stocks. Another factor favoring crude oil over petroleum product stocks is the greater flexibility in product mixes afforded by crude oil stocks. For example, heating oil is likely be relatively more valuable during disruptions that occur during the Northern Hemisphere's heating season, while gasoline is likely to be relatively more valuable during its summer months. Refineries can alter the product mixes they obtain from crude oil, thus making more of the products that are relatively highly valued at the time.

6. TIME AND RISK

Strategic stocks share several characteristics with insurance. Just as the insurance premium must be paid in advance of the coverage obtained, petroleum must be accumulated prior to its use to counter the economic losses of an oil price shock. Just as events justifying claims against insurance coverage may never materialize, strategic stocks may never be drawn down. Analyses of the desirability of strategic petroleum reserves require attention to the timing of costs and benefits as well as to uncertainty about whether benefits will actually be realized.

Valuing costs and benefits that accrue at different times requires that they be put into a common metric. Economists do so by converting future costs

and benefits to their equivalent present values. The present value recognizes that consumers generally prefer to consume sooner rather than later. Although there is considerable controversy over the appropriate social discount rate that should be used to convert future costs and benefits to their present values, the general procedure is straightforward. In the case of assessing strategic stockpiling programs, current costs required to accumulate stocks receive higher weights than benefits that accrue in the future.

The more challenging analytical task in assessing stockpiling programs is how to deal with uncertainty. Will an oil supply disruption warranting use of strategic stocks ever occur? If so, when will it occur? How severe will it be? How long will it last? None of these questions can be answered with certainty.

Analysts typically replace these uncertainties with corresponding risks. For example, assume that there is some probability that a major disruption will occur each year. They can then assess the net benefits of a particular stockpiling program for different assumed values of this probability.

The earliest approaches posited a representative disruption scenario occurring, say, 10 years after the stockpiling program begins. The present value of the costs of accumulating and storing the stocks were estimated, as were the present value of benefits that would result from draw down of the stocks during the disruption scenario. The question could then be asked: How high a probability of the scenario occurring would be required to yield an expected value of net benefits of zero, so that the stockpiling program would just break even? For example, imagine that the present value of the costs of accumulating and storing the stocks were estimated to be \$10 billion and the present value of the benefits were \$100 billion, then the probability of the posited disruption would have to be at least 10% for the stockpiling program to break even. Decision makers could then be asked if their subjective assessment of the probability of the representative scenario were greater than this breakeven probability.

More sophisticated treatments of uncertainty require models that allow for differences in the timing, magnitude, and duration of oil accumulation, supply disruptions, and draw downs. Such models typically posit a transition matrix that specifies assumed probabilities of moving from one market state to another, say moving among normal, minor disruptions, and major disruptions. Dynamic programming or other solution techniques are then used to find the optimal stock acquisition or draw downs for any particular combination of current market

state and stock size. In the United States, this approach generally supported expansion of the public reserves as well as large draw downs at the onset of major supply disruptions, even when there was a risk that the disruptions would persist beyond the point when stocks would be exhausted.

7. PROGRAMMATIC APPROACHES

OECD members have adopted a variety of approaches for meeting the International Energy Program requirement for holding 90 days of imports in stocks. All net importing members either require their domestic companies to hold strategic stocks or have standby authority to do so. Overall, company stocks account for approximately two-thirds of stocks held under the International Energy Program. In a number of countries, however, the stocks held in compliance with the International Energy program are largely working stocks that could not be fully deployed as strategic stock draw downs.

The most certain strategic stocks are those under direct government control in Japan, Germany, and the United States. Together, they could draw down 8.7 mmb/d during the first month of a major oil supply disruption. Such a rapid initiation of a coordinated draw down by these three countries at the outset of a supply disruption could substantially reduce its economic costs.

7.1 Japan National Oil Corporation

Even prior to the International Energy Program, Japan began requiring its firms to hold petroleum stocks. In 1978, it created the Japan National Oil Corporation. Currently, private firms have access to low-interest loans to cover the 25 days of imports they are required to hold above working stocks (45 days). The Japan National Oil Corporation has a goal of holding 90 days of imports. Currently, it has about 315 mmb in storage at 10 sites. Storage methods include conventional steel tanks, floating barges, and underground caverns.

The Ministry of Economy, Trade and Industry, subject to Cabinet decision, makes decisions about use of the stocks. Draw downs from company stocks are initiated by reducing the number of days of required stocks, as was the case in 1991 during the Gulf War; draw downs from the Japan National Oil Corporation would be directly ordered. Currently, the Japan National Oil Company could release 2.3 mmb/d during the first month of an all out draw

down, followed by rapid declines in succeeding months with negligible volumes by the sixth month.

7.2 Erdölbevorrantungsverband

The Oil Stockpiling Law of 1978 established the Erdölbevorrantungsverband (EBV) as an organization to hold petroleum stocks for Germany. All companies, except those in the chemical industry, that refine or import petroleum are compulsory members of the EBV. The EBV is funded solely through membership dues paid by companies in proportion to the quantities of gasoline, middle distillates, and fuel oils produced or imported. The dues are shown as itemized charges to retail customers. Thus, consumers rather than taxpayers finance the EBV. The EBV directors are appointed by an advisory board consisting of government and industry officials. Draw downs are initiated by the government under the Oil Stockpiling Law. Companies that have made higher dues payments receive higher priority in allocation.

In 1998, the government raised the stockholding requirement to 90 days of imports. Currently, the EBV stocks consist of 60% petroleum products and 40% crude oil, though the EBV has announced plans to reverse this proportion in order to increase flexibility in the products available during draw downs and reduce storage costs. Stocks are stored both regionally and in salt domes in Northern Germany. The EBV currently holds products equivalent to approximately 200 mmb of crude oil. During the first 3 months of a supply disruption, the EBV could support a draw down rate of approximately 2.3 mmb/d (0.9 mmb/d of crude oil and 1.4 mmb/d of products). A full draw down would exhaust reserves in 5 months.

7.3 The U.S. Strategic Petroleum Reserve

The Energy Policy and Conservation Act of 1975 called for the creation of a U.S. Strategic Petroleum Reserve of at least 150 mmb within 3 years and 500 mmb, then about 90 days of imports, within 7 years. Subsequent amendments increased the possible size to 1000 mmb. The Strategic Petroleum Reserve has capacity for holding 700 mmb of crude oil in salt dome caverns in Louisiana and Texas; it held 610 mmb crude oil in mid-2003. Since its permanent establishment in 2000, the Northeast Heating Oil Reserve has stored 2 mmb of heating oil in New Jersey, Connecticut, and Rhode Island.

The Strategic Petroleum Reserve provides substantial surge capacity. An all-out draw down of

4.1 mmb/d of crude oil could be sustained for approximately 3 months, with the rate declining in subsequent months as salt dome caverns are emptied. Routine draw downs occur on a regular basis to complete swaps with companies to adjust the grades of crude oil in storage. Draw downs have also been undertaken to raise revenue for deficit reduction. The president may order full-scale draw downs to counter severe energy supply interruptions that may cause major adverse impacts on national security or the national economy. Since 1990, the president has been authorized to initiate limited draw downs of up to 30 mmb in less severe circumstances.

The Strategic Petroleum Reserve was a draw down under presidential directive during the First Gulf War in cooperation with other International Energy Program countries. A week after the Iraqi invasion of Kuwait in August 1990, the secretary of energy announced a 5 mmb draw down to test distribution procedures and demonstrate resolve to use the reserve to counter an oil price shock. When the allied coalition air campaign began on January 16, 1991, the president put 30 mmb out to bid, 17 mmb of which were actually sold during the following 3 months.

The crude oil price fell almost \$10/b immediately after the announcement of the draw down, but analysts disagree about whether the price decline resulted primarily from anticipation of the draw down or changing perceptions about the risks to Gulf oil supplies. Analysts also disagree about whether sooner use of the reserve would have prevented the run up in prices during the latter part of 1990's that slowed economic growth.

The history of the Strategic Petroleum Reserve often showed a divergence between congressional authorization to expand stocks and the willingness of Congress and the administration to appropriate funds to enable it. Quite frequently oil purchases for the reserve were suspended for budgetary reasons, and steps have not been taken to implement the full 1000 mmb capacity. The use of in-kind royalties from federal offshore oil leases and heightened national security concerns in the wake of 9/11, however, make it likely that, absent an oil supply disruption, the reserve will be filled to its capacity of 700 mmb over the next several years.

8. CONCLUSIONS

Strategic petroleum reserves reduce the vulnerability of the world economy to oil price shocks. Members

of the International Energy Program have accumulated substantial strategic stocks, even discounting those that are actually working stocks held by companies and unlikely to contribute to surge capacity. These stocks will become more valuable as the proportion of world supply originating in the Persian Gulf increases. International and national policies appear to be continuing to evolve in several ways that increase the chances that strategic petroleum reserves will be used effectively.

First, policy analysts and policymakers now generally realize that strategic stocks have their primary effect through the world oil market. The draw downs of national stocks have very little capability for differentially benefiting those drawing down. Domestic oil prices below world levels simply divert imports to the other countries, thereby equalizing prices. Rather than viewing strategic stocks as providing national benefits in terms of relatively lower domestic prices, they should be seen as contributing to lower world prices, which in turn reduces the negative impacts of supply disruptions on the world economy.

Second, this realization, along with increasing economic globalization, should encourage further efforts toward early and internationally coordinated use of strategic stocks. It should also make individual countries with large surge capacities, namely the United States, Germany, and Japan, less hesitant to initiate draw downs even when agreement on coordination cannot be reached—any one country can realize large revenues from draw downs. Even if a country exhausts its strategic stocks, it will share in the benefits of later draw downs by other countries.

Third, although the language of the stockpiling agreement continues to emphasize total quantities of strategic stocks held, attention is increasingly being appropriately focused on maximum draw down rates during the first months of an oil supply disruption. This change in focus recognizes that a reduction in world supply has its most severe economic effects during the first 3 to 6 months before nondisrupted oil producers and consumers have time to change their mix of capital goods to in response to the higher prices.

Fourth, most national plans for draw downs take advantage of existing commercial networks. Allow-

ing markets to move petroleum to its most valued economic uses helps ensure that this especially scarce resource during supply disruptions will reach its most valued uses and thereby limit economic losses to the greatest extent possible.

SEE ALSO THE FOLLOWING ARTICLES

Geopolitics of Energy • Markets for Petroleum • National Energy Policy: United States • Nationalism and Oil • National Security and Energy • Oil and Natural Gas Liquids: Global Magnitude and Distribution • Oil Crises, Historical Perspective • Oil Price Volatility • OPEC, History of • Petroleum Property Valuation • Storage of Energy, Overview

Further Reading

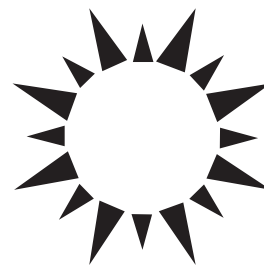
- Balke, N. S., *et al.* (2002). Oil price shocks and the U. S. economy: Where does the asymmetry originate? *Energy J.* 23(1), 27–52.
- Bohi, D. R., and Toman, M. A. (1995). Oil and national security: An assessment of externalities and policies. In “The New Global Oil Market: Understanding Energy Issues in the World Economy” (S. Shojai, Ed.), pp. 203–206. Praeger, Westport, CT.
- Horwich, G., and Weimer, D. L. (1984). “Oil Price Shocks, Market Response, and Contingency Planning.” American Enterprise Institute for Public Policy Research, Washington, DC.
- Horwich, G., and Weimer, D. L. (eds.). (1988). “Responding to International Oil Crises.” American Enterprise Institute for Public Policy Research, Washington, DC.
- International Energy Agency (2001). “Oil Supply Security: The Emergency Response Potential of IEA Countries in 2000.” Organization for Economic Development and Cooperation (OECD)/International Energy Agency (IEA), Paris, France.
- Kuenne, R. E., *et al.* (1975). A policy to protect the U. S. against oil embargoes. *Policy Analysis* 1(4), 571–597.
- Mead, W. J., and Sorensen, P. E. (1971). A national defense petroleum reserve alternative to oil import quotas. *Land Economics* 47(3), 211–224.
- Teisberg, T. J. (1981). A dynamic programming model of the U.S. strategic petroleum reserve. *Bell J. Economics* 12(2), 526–546.
- Toman, M. A. (1993). The economics of energy security: Theory, evidence, policy. In “Handbook of Natural Resource and Energy Economics 3: Economics of Energy and Minerals” (A. V. Kneese and J. L. Sweeney, Eds.), pp. 1167–1218. North Holland, Amsterdam.
- Verleger, P. K. (1990). “Oil Markets in Turmoil.” Ballinger, Cambridge, MA.
- Weimer, D. L. (1982). “The Strategic Petroleum Reserve: Planning Implementation and Analysis.” Greenwood Press, Westport, CT.

Subsidies to Energy Industries

DOUG KOPLOW

Earth Track Inc.

Cambridge, Massachusetts, United States



-
1. Introduction
 2. Subsidy Definition and Measurement
 3. Subsidies through the Fuel Cycle
 4. Patterns of Subsidization
 5. Subsidy Reform
 6. Conclusion

Glossary

consumer-subsidy equivalent Integrated metric of aggregate support provided to consumers by varied government policies; policies that act as taxes are incorporated by using the opposite sign.

cross-subsidies Policies that reduce costs to particular types of customers, products, or regions by increasing charges to other groups.

intermediation value Difference between the break-even cost of debt, insurance, or other programs to large governments and what these same services would cost if a smaller, higher risk private firm or individual had to buy them directly.

producer-subsidy equivalent Integrated metric of aggregate support provided to producers by varied government policies; policies that act as taxes are incorporated by using the opposite sign.

renewable portfolio standards Requirements mandating purchase of preset percentages of renewable electricity in particular service regions; the standards normally compete eligible supply sources against each other to minimize the per-unit subsidy.

subsidies Government-provided goods or services that would otherwise have to be purchased in the market or special exemptions from standard required payments or regulations; subsidies may be in cash but often involve shifting risks from private parties to taxpayers or the public.

Whether by intent or by accident, government interventions affect the relative prices of various energy technologies and with them the pattern of energy use

and investment. Depending on their structure, interventions can act as either taxes or subsidies to particular technologies, producers, or consumers. The lines can blur; some “taxes” may be insufficient even to cover the cost of beneficial government services, leaving a residual subsidy to the “taxed” parties.

1. INTRODUCTION

Energy resources vary widely in terms of their capital intensity, reliance on centralized networks, environmental impacts, and energy security profiles. Although the policies of greatest import to a particular energy option may differ, their aggregate impact may be significant. Subsidies to conventional fuels can slow research into emerging technologies, thereby delaying their commercialization. Subsidies and exemptions to polluting fuels reduce the incentive to develop and deploy cleaner alternatives. Inadequate tracking and recovery of costs associated with protecting energy security reduces the drive toward diversification.

Justifications for energy subsidies include social welfare, protection and promotion of jobs or industries, rural development, and energy security. Existing policies often fail to achieve these aims in practice. Because subsidies can be worth millions of dollars and often require sophistication or connections to obtain, policies implemented to help poorer segments of society may end up enhancing the wealth of more powerful groups instead. Often, the objectives of subsidy programs can be achieved in a manner that is more narrowly targeted and efficient than the subsidy policies now in place. For example, decoupling subsidy payments from resource-depleting activities can greatly reduce the environmental damages associated with the transfers.

This article examines the general issue of subsidy definition and measurement and then presents central issues associated with subsidization at each stage of a

generic fuel cycle. Aggregate patterns of subsidization and the challenges of subsidy reform are addressed in the subsequent two sections.

2. SUBSIDY DEFINITION AND MEASUREMENT

Government interventions encompass a wide range of regulatory, fiscal, tax, indemnification, and legal actions. By modifying the rights and responsibilities of various parties involved with the energy sector, these actions decrease (subsidize) or increase (tax) either energy prices or production costs. As discussed in what follows, differing approaches to subsidy definition and measurement have historically led to disparate subsidy analyses that are difficult to compare or compile.

2.1 Common Disagreements in Subsidy Definition

Disagreements over the proper definition of subsidy are common. Conflicts often arise over the form and timing of the transfer, the definition of the “non-subsidized” baseline, and the boundaries of analysis.

2.1.1 Form and Timing of Transfer

Energy subsidies are often viewed primarily as cash payments from a government agency to private businesses or individuals. Payments to low-income households to enable them to purchase heating oil and grants to businesses to help them develop particular energy technologies are examples. In reality, subsidies can take many different forms, and a more accurate definition must include government-provided goods or services, including risk bearing, that would otherwise have to be purchased in the market. Much market activity involves controlling and sharing the risks and rewards of economic activities, and risk-oriented subsidies are quite important. Subsidies can also be in the form of special exemptions from standard required payments such as tax breaks. Although cash payments are easily measured within a single year, more complex subsidies such as loan guarantees are best evaluated over multiple years so that patterns in losses or investment distortion can be seen more clearly.

2.1.2 Defining the Baseline

Subsidies must often be measured against a baseline. What would taxes owed have been in the absence of

this special tax break? How much would industry have had to pay in interest to build that new facility if the government had not guaranteed the loan? Many disagreements over subsidy definition originate in differing views on the appropriate baseline:

2.1.2.1 Indirect Versus Direct Transfers Some argue that interventions “count” as energy subsidies only if they directly target the energy sector. For example, the U.S. Energy Information Administration did not count subsidies to energy facilities provided by tax-exempt general-purpose municipal bonds in its tallies, arguing that the bonds did not constitute an energy subsidy if hospital and road construction could also use them. Similarly, fees to use the U.S. inland waterway system have historically been insufficient to reimburse system construction and maintenance costs. Although oil and coal industries are among the largest users of the system, the fee subsidies (which allow oil and coal a lower delivery cost than would otherwise be possible) are often discounted on the grounds that many commodities use the system. Because many subsidies tilt the energy-playing field toward a particular fuel even if there are also nonenergy beneficiaries, such policies should not be ignored. Rather, any policy that has the effect of subsidizing prices or production costs should be assessed. This may include policies targeted at single sectors, multiple sectors, specific geographic areas, or specific factors of production.

2.1.2.2 Externalities Although the levels vary by fuel, most energy production and consumption generate wide-ranging externalities such as pollution. Exemptions from appropriate environmental controls (e.g., less stringent air pollution control requirements for certain old power plants in the United States) penalize cleaner energy types and are properly viewed as subsidies. However, externalities do create an analytic challenge because they are difficult to monetize. Koplow and Dernbach documented that decisions about which externalities to include and how to value them can generate very large variation in subsidy estimates, on the order of hundreds of billions of dollars per year. Segregating externality-related subsidies from fiscal subsidies can help to improve data comparability and transparency.

2.1.2.3 Intermediation Value Subsidy recipients often incorrectly claim that risk-based subsidies such as loan guarantees have no value unless there are defaults. In addition to the direct cost of the government, all of these policies have an

intermediation value as well given that the government's cost of debt or indemnification is lower than what could be attained by the recipient on its own. For particularly high-risk endeavors, such as sales of large energy assets to politically unstable countries, the intermediation value can be very large. Because some energy resources have much greater access to these government programs than do others, distortions in relative energy prices can result. For example, transactional data on subsidized international lending assembled for the World Commission on Dams show that lending has heavily favored fossil fuels over renewables.

2.1.3 Boundaries of Analysis

Energy is a primary material, an input to refined energy products (e.g., gasoline, electricity), and an input to nonenergy materials (e.g., metals, consumer goods). To provide an accurate picture of energy subsidies, analytic boundaries with respect to three areas must be addressed: calculation of net values, level of government, and subsidies to complements.

2.1.3.1 Calculation of Net Values Because interventions can act as taxes or subsidies, interventions should be treated holistically so that end values represent net, rather than gross, subsidies to energy. Policies that affect multiple sectors need to be prorated so that only the portion applicable to energy is counted. Although proration is not always possible to do precisely (some policies have joint effects), reasonable allocations based on intensity of use, share of production, or similar metrics often provide reasonable proxies. Similarly, taxes or fees levied on fuels should be counted against the gross subsidy to generate the value net of offsets.

2.1.3.2 Level of Government Interventions occur at multiple levels of government, and all levels can affect energy costs to some degree. Analysis of these policies should be internally consistent; if tax offsets at the state level are deducted from energy subsidy values, state-level subsidies should be included as well.

2.1.3.3 Subsidies to Complements Energy subsidies have many second-order effects as they flow through other activities in the economy. Although detailed assessment is beyond the scope of this article, these policies can influence basic aspects of economic structure such as materials production, recycling, and energy demand. Subsidized electricity to primary aluminum is endemic and inhibits

adoption of less energy-intensive materials. Similarly, subsidies for converting wastepaper, animal waste, and landfill gas into energy reduce the viability of recycling and composting alternatives. Widespread subsidies to driving spur increased road use and demand for gasoline.

2.2 Methods of Transferring Value

Table I provides an overview of intervention types. Depending on the policy specifics, many interventions generate either a net subsidy or a net tax. Cash transfers from government to private industry may originate from a handful of the intervention types, including direct spending, government ownership, and research and development (R&D) support. However, subsidies resulting from avoided expenditures by private firms are more common. These include government provision of market-related information; access to below-market credit, insurance, or government-provided goods and services; tax breaks; and exemptions from prudent regulation on health-, environment-, or safety-related aspects of an enterprise.

Market controls, including general regulations, provisions governing access to resources, restrictions on energy-related imports or exports, and purchase requirements can act as either a tax or a subsidy depending on one's market position. For example, past restrictions on oil exports from Alaska acted as a *de facto* tax on Alaskan producers because they could not sell output to the highest bidder. However, the very same policy provided oil subsidies to consumers on the West Coast of the United States. Although driven by government policy and having important effects on energy market structure, these policies often involve transfers between various producers and consumers rather than transfers from taxpayers. Cross-subsidies follow a similar pattern, with some users paying less than they should and others paying more than they should. They commonly occur when rate structures must be approved by governments or in markets protected from competition.

Special energy taxes, as their name suggests, tend to act as taxes rather than as subsidies. However, a tax should be classified as "special" only if it is above and beyond appropriate baseline recovery of revenue. The baseline taxation of energy should (1) compensate public sector owners for the sale of valuable energy resources, (2) recover public sector costs associated with the public provision of energy-related services, (3) equal the baseline tax on

TABLE 1

Common Forms of Government Interventions in Energy Markets

Intervention type	Description
Access ^a	Policies governing the terms of access to domestic onshore and offshore resources (e.g., leasing)
Cross-subsidy ^{a,b}	Policies that reduce costs to particular types of customers or regions by increasing charges to other customers or regions
Direct spending ^b	Direct budgetary outlays for an energy-related purpose
Government ownership ^b	Government ownership of all or a significant part of an energy enterprise or a supporting service organization
Import/Export restriction ^a	Restrictions on the free market flow of energy products and services between countries
Information ^b	Provision of market-related information that would otherwise have to be purchased by private market participants
Lending ^b	Below-market provision of loans or loan guarantees for energy-related activities
Price controls ^a	Direct regulation of wholesale or retail energy prices
Purchase requirements ^a	Required purchase of particular energy commodities, such as domestic coal, regardless of whether other choices are more economically attractive
Research and development ^b	Partial or full government funding for energy-related research and development
Regulation ^a	Government regulatory efforts that substantially alter the rights and responsibilities of various parties in energy markets or that exempt certain parties from those changes
Risk ^b	Government-provided insurance or indemnification at below-market prices
Tax ^{a,b}	Special tax levies or exemptions for energy-related activities

Source: Koplow, D. (1998). "Quantifying Impediments to Fossil Fuel Trade: An Overview of Major Producing and Consuming Nations." Paper prepared for the OECD Trade Directorate.

^aCan act either as a subsidy or a tax depending on program specifics and one's position in the marketplace.

^bInterventions included within the realm of fiscal subsidies.

other commodities, and (4) charge an appropriate levy for negative externalities associated with production and use of the resource. Many assessments of energy taxation fail to incorporate appropriate measures for baseline levels, improperly classifying as an "energy tax" or an "environmental tax" policies that in fact recover only a portion of public costs and leave residual subsidies.

Government ownership of energy-related enterprises, including power generation and transmission, oil production and refining, coal mines, and road and pipeline networks, is common around the world. Many of these activities generate large subsidies to consumers as well as depleting fiscal resources. Subsidies are multilayered. The enterprise is often fairly high risk, attains access to low-cost tax-exempt government debt, pays no taxes on net income, and is not expected to earn a return on capital. Operating losses may ensue above and beyond these cost structure subsidies due to poor controls or politicization of the rate structure. These enterprises can be complex and difficult to analyze but often contribute to significant energy market distortions.

2.3 Methods of Measuring Subsidy Magnitude

Efforts to assess subsidy magnitude have generally focused either on measuring the value transferred to market participants from particular programs (program-specific approach) or on measuring the variance between the observed and the "free market" price for an energy commodity (price gap approach). One set of methods that captures both pricing distortions (net market transfers) and transfers that do not affect end market prices (net budgetary transfers) is the producer subsidy equivalent (PSE) and consumer subsidy equivalent (CSE) metrics commonly employed in the agricultural sector. Use of these in the energy sector has so far been limited to annual assessments of PSEs for coal in a handful of countries.

These approaches differ in the amount of data required to calculate them and in the degree to which they successfully measure budget transfers plus market transfers. Program-specific transfer assessments capture the value of government programs benefiting (or taxing) a particular sector, whether these benefits end up with consumers (as lower

TABLE II
Overview of Subsidy Measurement Approaches

Approach/description	Strengths	Limitations
<i>Program specific:</i> Quantities value of specific government programs to particular industries; aggregates programs into overall level of support	Captures transfers whether or not they affect end market prices Can capture intermediation value (which is higher than the direct cost of government lending and insurance)	Does not address questions of ultimate incidence or pricing distortions Sensitive to decisions on what programs to include Requires program-level data
<i>Price gap:</i> Evaluates positive or negative “gaps” between the domestic price of energy and the delivered price of comparable products from abroad	Can be estimated with relatively little data; very useful for multicountry studies Good indicator of pricing and trade distortions	Sensitive to assumptions regarding “free market” and transport prices Understates full value of supports because ignores transfers that do not affect end market prices
<i>PSE/CSE:</i> Systematic method to aggregate transfers plus market supports to particular industries	Integrates transfers with market supports into holistic measurement of support Separates effects on producer and consumer markets	Data intensive Little empirical PSE/CSE data for fossil fuel markets

Source: Koplow, D., and Dernbach, J. (2001). Federal fossil fuel subsidies and greenhouse gas emissions: A case study of increasing transparency for fiscal policy, *Annu. Rev. Energy Environ.* 26, 361–389.

prices), producers (through higher revenues), or resource owners (through higher rents). Unless integrated into a macroeconomic model, this information tells little about the ultimate incidence of the subsidy programs and their effect on market prices. By definition, the price gap metric highlights observed price distortions, although it misses the often substantial fiscal supports that do not affect consumer energy prices but do affect the structure of supply. The combination of PSE and CSE data provides insights into both. Table II briefly summarizes the main approaches that have been used in both domestic and international subsidy assessments as well as their respective strengths and limitations.

3. SUBSIDIES THROUGH THE FUEL CYCLE

Because no two fuel cycles are exactly the same, examining subsidies through the context of a generic fuel cycle is instructive in providing an overall framework from which to understand how common subsidization policies work. Subsidies are grouped into preproduction (e.g., R&D, resource location), production (e.g., extraction, conversion/generation, distribution, accident risks), consumption, postproduction (e.g., decommissioning, reclamation), and externalities (e.g., energy security, environmental, health and safety).

3.1 Preproduction

Preproduction activities include research into new technologies, improving existing technologies, and market assessments to identify the location and quality of energy resources.

3.1.1 Research and Development

R&D subsidies to energy are common worldwide, generally through government-funded research or tax breaks. Proponents of R&D subsidies argue that because a portion of the financial returns from successful innovations cannot be captured by the innovator, the private sector will spend less than is appropriate given the aggregate returns to society. Empirical data assembled by Margolis and Kammen supported this claim, suggesting average social returns on R&D of 50% versus private returns of only 20 to 30%.

However, the general concept masks several potential concerns regarding energy R&D. First, ideas near commercialization have much lower spillover than does basic research, making subsidies harder to justify. Second, politics is often an important factor in R&D choices, especially regarding how the research plans are structured and the support for follow-on funding for existing projects.

Allocation bias is also a concern. Historical data on energy R&D (Table III) demonstrate that R&D spending has heavily favored nuclear and fossil energy across many countries. Although efficiency,

TABLE III

Federal Research and Development Support by Energy Type

Region	Nuclear fission	Fossil fusion	Fossil energy	Renewables, efficiency, conservation	Other ^a
United States (percentages)					
1950–1993 ^b	49.2	13.1	21.5	16.2	0
1998–2003 ^c	17.7	16.2	32.6	30.2	3.4
IEA members ^d (percentages)					
1974–1993	49.7	10.8	14.1	14.0	11.4
1994–1998	39.3	11.0	10.6	20.8	18.4
Total spending (billions of 2001 dollars)					
1974–1998	117.3	26.5	33.0	36.8	30.6
			Coal: 23.5	Renewables: 19.8	
			Oil and gas: 9.5	Conservation: 17.0	

^aIncludes electrical conversion and distribution, energy storage, and unclassified spending.^bKoplow, D. (1993). "Federal Energy Subsidies: Energy, Environmental, and Fiscal Impacts: Technical Appendix." Alliance to Save Energy, Washington, DC.^cU.S. Department of Energy. (2003). "Budget Authority History Table by Appropriations."^dInternational Energy Agency. (2003). *Research and Development Database*.

renewables, and conservation have captured a higher share of public funds during recent years, the overall support remains skewed to a degree that may well have influenced the relative competitiveness of energy technologies. Extensive public support for energy R&D may also reduce the incentive for firms to invest themselves. U.S. company spending on R&D for the petroleum refining and extraction sector was roughly one-third the multi-industry average during the 1956–1998 period based on survey data from the U.S. National Science Foundation. For the electric, gas, and sanitary services sector, the value was one-twentieth, albeit during the more limited 1995–1998 period.

3.1.2 Resource Location

Governments frequently conduct surveys to identify the location and composition of energy resources. Although these have addressed wind or geothermal resources on occasion, they most often involve oil and gas. Plant siting is another area where public funds are used, primarily to assess risks from natural disasters such as earthquakes for large hydroelectric or nuclear installations. Survey information can be important to evaluate energy security risks and to support mineral leasing auctions, especially when bidders do not operate competitively. However, costs

should be offset from lease sale revenues when evaluating the public return on these sales. Similarly, the costs of siting studies should be recovered from the beneficiary industries.

3.2 Production

Energy production includes all stages from the point of resource location through distribution to the final consumers. Specific items examined here include resource extraction, resource conversion (including electricity), the various distribution links to bring the energy resource to the point of final use, and accident risks.

3.2.1 Extraction of Energy Resources

General procedures for leasing access to energy minerals on public lands and more general subsidies for promoting energy extraction both are important areas to evaluate. Extraction-related subsidies are most common for oil and gas production, although they also support nuclear fission (due to uranium mining), geothermal, and coal.

3.2.1.1 Accessing Publicly Owned Energy Resources Terms of access for energy minerals on public lands can be a source of enormous subsidies. In

countries where leases or concessions are granted through graft rather than competitive bidding, wealth transfers worth billions of dollars can occur. Although there are not good statistics on the losses, the problem appears to be large. Oxfam America finds that states most dependent on oil tend to have very low Human Development Index (HDI) rankings. The HDI, developed by the UN Development Program, ranks states according to a combined measure of income, health, and education. Transparency International finds strong linkages between large mining and petroleum sectors as well as elevated levels of bribery and corruption. Low-cost access to energy minerals also tends to remove the incentive for careful management because profits can be had even with inefficient operation. Lease operation can also generate subsidies such as when self-reported royalties are calculated improperly. The Project on Government Oversight has documented state and federal court awards in excess of \$10 billion in response to litigation in the United States over oil and gas royalty underpayments.

3.2.1.2 Promoting Extraction Activities Policies to reduce the cost of extraction are widespread and often take the form of tax or loan subsidies or royalty concessions. They are found at both the national and state levels. Particular market niches may be targeted, from geographical (e.g., deep sea recovery of oil, timbering in a particular forest), to technological (e.g., tax breaks for more advanced oil drilling or coal gasification), to life cycle related (e.g., lower royalties on idle wells that are restarted). In some cases, baseline tax policy may be applied by firms in creative ways to generate large subsidies. U.S.-based multinationals receive a tax credit for foreign taxes paid to avoid double taxation of foreign income. Yet in many oil-producing regions with low or no corporate income taxes, foreign governments have reclassified royalty payments into corporate taxes, generating a tax write-off estimated by Koplow and Martin at between \$0.5 billion and \$1.1 billion annually.

However, many subsidies to extraction are not restricted to particular market niches. Percentage depletion allowances in the United States allow most firms mining oil, gas, uranium, or coal to deduct more costs from their taxable income than they have actually incurred. Accelerated write-offs of extraction-related investments are also common. For example, many multiyear costs in the U.S. oil and gas industry may be deducted immediately (expensed) rather than over the useful lives of the investments. All of these special provisions tend to

reduce the effective tax rate on benefiting energy industries. Data collected by the Energy Information Administration (EIA) suggest that the major U.S. energy firms paid federal taxes that were one-quarter to one-half the prevailing nominal rates between 1977 and 1995.

3.2.2 Conversion

Raw energy materials normally go through some conversion prior to consumption. Crude oil is refined into a wide range of specialized products such as gasoline and heating oil. Coal may be pulverized or cleaned prior to use. A combination of heat and machinery converts raw fuels (including wind and solar) into electricity. Common government supports to the conversion stage include capital subsidies, production tax credits or purchase requirements, and exemptions from appropriate protections for environmental quality, worker health, and accident risks. Because this third category affects multiple phases of the fuel cycle, it is addressed in a separate section.

3.2.2.1 Capital Subsidies Subsidies to capital formation, usually through accelerated depreciation or investment tax credits, are common. Although applicable to multiple economic sectors, they are often of great benefit to energy producers. This is due both to their relative capital intensity and to provisions in the tax code that grant special accelerated depreciation schedules for energy-related assets. For example, in the United States, three sectors of relevance to energy—electric light and power, gas facilities, and mining, shafts, and wells—have allowable depreciation schedules that are 28, 45, and 44% faster, respectively, than the actual economic depreciation of their assets according to data compiled by the U.S. Treasury. Capital subsidies are of greatest benefit to large-scale generation assets with long construction times (nuclear, hydro, and coal) and are of greatest detriment to energy resources that conserve capital (most prominently energy conservation).

3.2.2.2 Tax Credits/Purchase Mandates A second class of subsidies to the conversion stage are tax credits or purchase mandates for certain types of energy. These subsidies occur at both the federal and the state/provincial levels and most often support emerging power sources such as solar, wind, and biomass-based electricity. Whereas many of the subsidies to conventional power sources are expensive regardless of whether the energy investments ultimately succeed, the tax credits and purchase mandates tend to be more efficient. For example,

federal tax credits for wind energy in the United States cost taxpayers nothing unless a private investor is successful in getting a wind plant operating. If the plant goes offline, so too do the credits. Renewable portfolio standards (RPSs), a common form of purchase mandates adopted by many U.S. states, are even more efficient. In addition to providing no subsidy unless the power is delivered, RPSs often compete eligible power sources against each other, driving down the unit subsidy as technologies improve.

Despite their benefits, these approaches have run into some political problems. Specifically, as the subsidies have grown, so too has lobbying pressure to expand the range of eligible sources. Federal tax credits now include poultry waste, a great benefit to the handful of very large chicken processors. At the state level, unsustainable biomass sources are sometimes included, as are waste-to-energy and landfill gas systems. Thus, although energy diversification

goals are still being met, the supply is not necessarily renewable or particularly clean.

3.2.3 Transportation and Distribution

Fuel cycles may involve multiple transportation steps, including movement of raw fuels to point of refining, refined fuels to the point of consumption, and movement of wastes to disposal sites. Relevant modes of transport include road, rail, water, pipelines, and transmission lines.

Although specific energy resources vary widely in their transport intensity and in the modes of transportation and distribution on which they rely (Table IV), there are some common themes. Government construction, maintenance, and operation of transportation infrastructure frequently give rise to subsidies when user fees do not cover costs. These subsidies are often understated because municipalities might not properly cost the resources being

TABLE IV
Impact of Transport Subsidies on the Energy Sector

Transport mode	Issues	Energy sector impacts
Water: Inland	Waterway maintenance often provided by governments; user fees may not recover costs	Reduces delivered price of bulk oil and coal
Water: Coastal and international	Coastal ports, harbors, and shipping oversight subsidized by federal and other government entities; user fees might not recover costs Fuel consumed during shipment in international waters generally tax free	Reduces delivered price of bulk oil and coal
Road	Most roadways are municipally owned and operated; user fees (primarily from fuel taxes) often insufficient to cover costs Large trucks often pay proportionately less in taxes than the damage they cause roadways	Primarily benefits refined petroleum products Waste products from coal combustion or waste-to-energy plants may sometimes move by truck as well
Rail	Many rail lines do not recover their full costs	Largest beneficiary is coal, with some benefits to oil
Pipeline	Rights of way, safety and security, and environmental cleanup contribute to reduced costs of pipeline ownership and operation Depending on circumstances, government ownership may generate large subsidies to users and use government monopoly to levy high taxes on users	Primarily benefits oil and natural gas
Electrical Transmission Grid	Rights of way, tax breaks for municipal ownership or capital investment, and government research and development can generate subsidies to electrical distribution Inaccurate pricing of distance can generate cross-subsidies to rural users	Benefits all sources of centralized electricity in proportion to their share as a prime mover in generating stations; coal, nuclear, natural gas, hydroelectricity, and oil are the main beneficiaries

Source: Earth Track Inc., Cambridge, MA.

consumed. For example, tax exemptions on transportation bonds used to finance roads are routinely ignored, as are the free grants of rights-of-ways for rail, road, pipeline, and transmission links. So too is the opportunity cost of land covered by roadways and parking facilities. Although this space occupies 1.7, 2.1, and 3.5% of the total land area in the United States, Germany, and Japan, respectively, Todd Litman of the Victoria Transport Policy Institute noted that no property tax is paid on the vast majority of this space. This understates the direct costs of the infrastructure and the rights to use it.

Cross-subsidies between user groups may further distort relative prices. Large trucks pay less in highway fees than the damage they cause, generating an incremental subsidy to deliveries of refined fuels such as gasoline. Deep-berth ships such as large oil tankers may be the primary drivers of channel- or port-deepening projects, yet they often contribute to costs based only on volume of shipments. In the electricity sector, transmission tariffs may represent broad averages of the cost of service rather than rising as the distance traveled and density of users decline. By delivering subsidized electricity to remote users, transmission cross-subsidies mask the cost of line maintenance and new construction. This can destroy niche markets in which off-grid technologies (often renewable) would otherwise have been able to compete. Cross-subsidies between peak pricing and low demand periods are also common in electricity markets because real-time metering is not widely used at the retail level. This can dampen retail investments in demand-side management.

Power sources such as wind and solar require no shipment of input fuels or waste. Improved energy efficiency and some off-grid technologies require no transmission grid either. As a result, subsidies to energy transport can increase the barriers to renewable energy and efficiency. A major U.S. study conducted by Cone and colleagues in 1978 found that an estimated \$15.2 billion in federal money subsidized transport of U.S. oil stocks between 1950 and 1977. The policies generating these subsidies have continued during the ensuing quarter-century or so.

3.2.4 Accident Risks

A handful of energy activities have the potential to cause catastrophic harm, including large oil spills, dam failures, and nuclear accidents. Many governments cap, shift, or ignore the potential liabilities from these activities. Functioning insurance markets and litigation would normally help to drive up prices for the more dangerous energy sources or particu-

larly negligent operators. Government policies that mask these signals impede substitution to safer alternatives.

3.2.4.1 Large Oil Spills Within the United States, the Oil Pollution Act of 1990 stipulates use of commercial insurance for a first tier of insurance. A public trust fund financed by levies on oil sales provides supplemental coverage, although payments out of the fund are capped at \$1 billion per incident. Based on empirical assessments of spill cleanup costs by Anderson and Talley, at least five spills over the past three decades or so would have exceeded the \$1 billion cap, although most spills will be adequately covered. Internationally, the 1992 Civil Liability Convention governs liability for oil spills, also using a two-tier system. Insurance held by the vessel owner provides the first tier. Levies on cargo owners feeds the second tier, with receipts held in the International Oil Pollution Compensation Fund. The maximum compensation available from both tiers is roughly \$174 million, a level shown to be insufficient by spills occurring in both 1997 and 1999. Although the caps are likely to be raised by 50%, Alfred Popp, chairman of the group working on the latest rounds of reforms, noted that concerns about liability shortfalls persist. The subsidy value of these caps is not known.

3.2.4.2 Dam Failures Many activities that would pose a very large potential risk if accident scenarios materialized rely on a system of strict liability. Strict liability focuses only on magnitude of the potential damages rather than on the intent, negligence, or degree of care of the operator. Although the failure of a large dam near a populated area can cause catastrophic loss of life, assurance for such potential liabilities is poorly characterized. Although loss of life from a dam failure will likely trigger widespread litigation, the rules of that litigation are predominantly set at the state level. Analysis by Denis Binder for the Association of State Dam Safety Officials indicates that a slight majority of states reject strict liability in dam failures. Furthermore, the piecemeal approach to coverage within the United States makes it difficult to evaluate whether existing coverages are adequate. Poor characterization of the risks extends to the international arena as well. To the extent that liability insurance is not in place or is too low, subsidies to hydroelectricity would result.

3.2.4.3 Nuclear Accidents Nuclear accidents can expose large populations to dangerous levels of radioactivity, triggering enormous liabilities for the

firm responsible. Caps on nuclear liability are common throughout the world. The United States, under the Price–Anderson Act, has a two-tier system of indemnification: a first tier of commercial insurance (\$300 million per reactor) plus a second pooled tier (maximum of \$83.9 million per reactor) funded by retroactive assessments on all reactors in case any reactor has an accident. Japanese nuclear operators must provide financial security of \$520 million; damages above that amount will be paid by the government. In China, the limit is roughly \$36 million. In Ukraine, it is roughly \$70 million.

International efforts to standardize liability under the Convention on Supplementary Compensation for Nuclear Damage would establish minimum liability coverage worldwide, although for many countries this would also constitute the maximum. Under the convention, operators would directly face a first tier of liability. A country fund would provide secondary coverage. Because country payments rely on a sovereign guaranty rather than a prefunded instrument such as a trust fund, they may be at some risk.

Aggregate coverage under the U.S. system is estimated at roughly \$9.2 billion per accident, although most of this is paid out over nearly 9 years by utilities, so the present value of the coverage is substantially lower. Liability levels established under the convention would provide less than \$900 million per accident. Loss statistics from the Insurance Services Office and from the Disaster Insurance Information Office provide some context. Since 1950, there have been approximately 20 hurricanes with adjusted damages in excess of the convention cap, and both Hurricane Andrew and the Northridge earthquake had damages that exceeded the Price–Anderson cap even before adjusting retroactive premiums to present values.

Subsidies arise when government caps fall below expected damages from an incident and caps under both Price–Anderson and the convention are likely to do so. Damages above that level are, in effect, shifted to the state or to the affected population. Heyes estimated that the subsidy to reactors under Price–Anderson ranges between 2 and 3 cents per kilowatt-hour, a value that would roughly double the operating costs of nuclear plants. In addition, there are incremental subsidies associated with indemnification for nuclear contractors and government-owned facilities. Because other countries have lower liability caps and weaker inspection regimes, they likely have higher liability subsidies as well.

3.3. CONSUMPTION

Government support for energy consumption falls into three main categories: poverty alleviation, economy-wide below-market pricing, and targeted subsidies for certain classes of consumers.

3.3.1 Poverty Alleviation

Subsidies to heat and power for poorer citizens are common, frequently in the form of a lump-sum grant or reduced cost access to municipal resources. Often consumption oriented, these subsidies may miss opportunities to implement conservation measures among the target populations. Targeting can be a problem as well, with funds not reaching the groups most in need. According to the International Energy Agency (IEA), the poorest citizens often rely on noncommercial fuels such as dung (biomass comprises as much as 80% of the energy market in rural countries with a high reliance on subsistence agriculture) or live outside the reach of the subsidized electrical grid.

3.3.2 General Subsidies

Nations with large domestic energy industries sometimes institute policies that keep local prices well below world levels. These subsidies may protect antiquated energy-consuming industries that otherwise would be unable to compete, or they may serve as “rewards” to the electorate for supporting a particular official. For example, price gap data for Venezuela and Iran compiled by the Organization for Economic Cooperation and Development (OECD) and IEA show that these large oil producers heavily subsidize both industrial and residential use of petroleum. Subsidies are also common in many service areas close to large municipal hydroelectric generating stations. For example, rates to customers of the Power Marketing Administration dams in the United States were long heavily subsidized. Although the quantities of power or oil flowing through these regions make these subsidies seem costless, they are not. Domestic sales at subsidized rates forgo energy export revenues, increase local pollution, and contribute to a production base that is increasingly noncompetitive with that deployed elsewhere in the world.

3.3.3 Targeted Exemptions

Most OECD countries exempt coal and heavy fuel oils used in industry, as well as aviation fuels used on international flights, from the baseline levies on energy. Excise tax rates on coal used in the industrial

or power sector are often lower than those on much cleaner natural gas. The OECD noted that these exemptions “effectively mean that a large proportion of total carbon emissions in OECD countries is untaxed,” generating weaker incentives to adopt even low-cost abatement options.

3.4 Postproduction Activities

Energy production and conversion require large facilities, often located in remote or pristine environments. Postoperational cleanup can be complex. Decommissioning addresses removal of physical infrastructure, whereas remediation and reclamation address problems with land and water. For markets to make accurate decisions about the relative cost of energy resources, the cost of these postproduction activities must be included in energy prices during the operating life of the facility in much the same way that the cost of an employee pension would be. Indeed, failure to accrue funds for postclosure costs during operations would make public subsidy likely given that revenues often drop to zero on plant closure.

3.4.1 Decommissioning

Decommissioning subsidies arise when infrastructure removal costs are ignored or underestimated or when accrued funds are mismanaged. Costs can be significant at large-scale energy installations such as hydroelectric dams and oil refineries. Where installations are remote (e.g., offshore oil rigs), radioactive (e.g., nuclear plants), or widely dispersed (e.g., gathering pipelines), costs per unit of capacity can be particularly high. Requirements for long-term environmental or safety monitoring (e.g., nuclear plants and some mines) can drive costs up further.

Pipelines and hydroelectric dams provide examples of costs being ignored entirely. Koplow and Martin made inquiries to many U.S. officials regarding pipeline closure. They found that although there are regulations governing proper abandonment, advance funding of closures was not required. The risks of insolvency appeared to be fairly high, especially for the smaller companies that often own older gathering pipelines. Regarding dams, the U.S. Federal Energy Regulatory Commission indicated in a 1994 policy statement that it will “not generically impose decommissioning funding requirements on licensees” but rather will stipulate them on a case-by-case basis at the time of relicensing. According to Andrew Fahlund of American Rivers, this policy has been implemented such that if a “dam owner is too

poor, it is too burdensome to require them to maintain a fund, and if they are rich, they will have plenty of money available for such an eventuality.”

Underestimating decommissioning requirements is of great concern with nuclear plants. IEA data indicate that the anticipated cost per unit of power capacity can vary by a factor of 10 across plants. IEA multicountry data suggest median decommissioning values of between 21 and 37% of the overnight capital cost (i.e., before financing) to build the plant. If funds are not properly accumulated during the plant’s operating life, taxpayer burdens will be large. Inadequate provision for closure is also apparent in the oil and gas sector. Koplow and Martin found shortfalls in funding to plug and abandon oil wells in the United States approaching \$600 million per year, of which approximately 75% represented insufficient bonding at wells still in operation.

Public bailouts can also be required if accrued funds for postclosure activities are lost through negligence, bankruptcy, or theft. If funds are retained within the firm, bankruptcy is a significant risk, especially given the 40- to 60-year time frame between fund collection and use. Increased segregation of each energy asset into its own company (now becoming the norm in the U.S. nuclear industry) greatly increases this risk. Loss through negligence is less likely where regulations preclude speculative investing. Nuclear decommissioning trusts within the United States are held outside the firm and are subject to conservative investment requirements to reduce the likelihood of loss.

3.4.2 Reclamation and Remediation

Small subsidies to site reclamation and remediation may arise through government-sponsored research into remediation technologies or through regulatory oversight of extraction activities that are not recovered via user fees. Much larger subsidies are associated with remediation of government-owned energy-related installations or where reclamation bonding has been insufficient to pay for the damage caused by private operators. James Boyd at Resources for the Future pointed to widespread inadequacy of reclamation bonding levels. For example, in the U.S. states of Indiana, Kentucky, and Tennessee, reclamation of coal mine sites is below 20%. Reclamation bond levels have generally been inadequate. Estimated liability for high priority (public health and safety concerns) coal mine remediation in the United States is \$6.6 billion, according to the U.S. Office of Surface Mining Reclamation and Enforcement. Many mining regions

around the world have unreliable, incomplete, or nonexistent data on abandoned mines and their associated costs. These shortfalls may be made up by general tax revenues. However, more often, resource damage is not mitigated and continuing environmental releases are not controlled. Spending to address environmental concerns at nuclear energy-related infrastructure owned by the U.S. government has run approximately \$500 million per year, much of which is paid by general taxpayers.

3.5 Energy Externalities

External costs of energy production and consumption can include pollution, land degradation, health impairments, congestion, and energy security. This article differentiates between two types of subsidies. The first involves existing government spending to address recognized problems associated with particular energy resources. Included here would be public funding to protect energy supplies and assets; public absorption of energy worker health care costs; and/or public subsidies to pollution control or abatement. Because this spending involves actual outlays, it is counted as a fiscal subsidy. A second class of policies involves loopholes in regulatory controls that allow additional damages to human health or the environment to continue without compensation. This second group is often difficult to quantify and is segregated as an externality.

3.5.1 Energy Security

Energy plays a central role in industrialized economies, and supply disruptions can trigger widespread economic dislocations. Geopolitical problems, accidents, and terrorism all are potential triggers. Lovins and Lovins identified a handful of factors that drive security concerns. These include long distribution channels, geographically concentrated delivery or production systems, interconnected systems that can spread failures, specialized labor and control systems to operate capital-intensive facilities that are very difficult to replace, and dangerous materials that can elevate the severity of any breach.

Energy security strategies include protection of energy-related assets and supply routes, stockpiling of vulnerable resources, and supply diversification. Where costs of these responses are borne by the general public rather than by the appropriate energy producers and/or consumers, the market incentive to build a more resilient, decentralized, and diversified supply system is reduced. Security subsidies tend to benefit oil the most, with particularly high transfers

to imported oil from unstable regions such as the Persian Gulf. Additional beneficiaries are centralized electricity and natural gas. Off-grid power and conservation are the sources most disadvantaged. Subsidies to protecting energy installations and stockpiling are explored in detail in the following subsections.

3.5.1.1 Protection of Assets and Supply Links The larger the energy installation, the greater the target and the bigger the dislocation that an attack or accident would cause.

Defending energy-related assets is an increasing concern of governments around the world. Pipeline defense is listed as its own objective within Georgia's defense and security strategy. The United States has become involved with training the Colombian military to defend oil pipelines in that country, pushing for funding of \$98 million to support the effort. Within the United States, core assets include the Trans-Alaska Pipeline System (TAPS), through which nearly 25% of total U.S. crude production flows, and nuclear plants. In response to inquiries from Koplow and Martin, Alaskan and federal officials said that no public funds were spent ensuring TAPS security. Nonetheless, the military has historically conducted training and planning exercises around the pipeline. In the nuclear sector, increased public subsidies have come through rising deployment of state-level security or National Guard troops around plants during periods of high terrorist alerts. However, surveys of nuclear plant workers by the Project on Government Oversight reveal employee concerns that training and spending levels are still insufficient. Although these anecdotes indicate that public expenditures in the area of protecting energy-related assets are likely large, data to quantify the subsidies are generally unavailable.

The costs of defending oil shipments through the Persian Gulf is an exception. As one of three central missions for the U.S. military in the region, there have been multiple efforts to value the subsidy to oil. Koplow and Martin reviewed eight historical studies of these costs and found general agreement that this presence is of great benefit to oil supply security. Disagreements centered on cost attribution. Some assessments attributed an extremely small portion of the military cost to oil, arguing that the same basic force structure would be needed for the other missions. Koplow and Martin pointed out that equivalent arguments could be made for each mission area given that the common costs of the vessels and personnel are what are most expensive.

They argued instead for treating the military presence through the lens of joint costs and allocating a reasonable portion (in this case, one-third) to the oil sector. This approach yields a subsidy to the oil sector in the range of \$11.1 billion to \$27.4 billion per year (roughly \$1.65–\$3.65/barrel originating from the region), depending on which of the detailed costing studies are used. Although funded by U.S. taxpayers, the benefits accrue to oil consumers in Europe and Japan as well. Recovering this cost via an excise fee on shipments would help to encourage increased supply diversification.

3.5.1.2 Stockpiling Petroleum has been the main focus of stockpiling efforts given its importance to world transport and military readiness. Under the terms of the IEA, oil-importing member states are required to hold stocks equal to 90 days of the previous year's net oil imports as a buffer against short-term supply disruptions. Subsidies arise if the costs of stockpiling are borne by taxpayers rather than by oil consumers. Relevant expenses include constructing and operating the stockpiles, interest costs on oil inventories and infrastructure, and any payments to third parties for nongovernmental stockpiling (two-thirds of IEA-mandated stocks are held commercially).

Buffer stocks for oil within the United States are held within the publicly owned Strategic Petroleum Reserve (SPR). The SPR has incomplete cost accounting, most prominently ignoring the interest costs associated with more than \$16 billion it has spent to purchase its oil inventory since the reserve's inception. Private firms must finance all working capital, including inventory, in their operations, and cost savings from reducing inventory levels can be large. Public oil stockpiles are no different; capital tied up in the enterprise must be borrowed, at interest, through Treasury bond markets. Analysis by Koplow and Martin for 1995 estimated annual subsidies to the SPR at between \$1.7 billion and \$6.1 billion, depending on whether unpaid interest on oil inventories is compounded. Because carrying costs are sensitive to the cost of capital, declining interest rates during recent years mean that current SPR subsidies will be lower than they were during the mid-1990s. Although the details of stockpile financing in other countries are not easy to discern (the IEA collects data only on physical flows, not on financial flows), some countries do recover at least a portion of their stockpiling costs from consumers. These include Japan, France, Germany, Korea, and Taiwan.

Subsidies to stockpiles slow transition to less vulnerable, more diversified supplies. Formal tracking of stockpile finance by the IEA, as well as the formalization of accounting rules for calculating costs, would leverage market forces for improved supply security.

3.5.2 Environmental, Health, and Safety Externalities

Externalities involve damages associated with energy production or use that are imposed on surrounding populations or ecosystems without compensation. These may include environmental damage, materials damage, human health effects, and nuisance factors such as bad smells and loud noises. Worker health is sometimes not counted as an externality under the argument that workers are compensated for the added risks of their jobs through higher wages. Such a conclusion requires that workers have some degree of choice in whether or not to accept jobs and that employers can be taken to task retroactively for gross negligence. This is not the case in many countries around the world. As a result, it is reasonable to consider as subsidies high levels of occupational illness, especially when the costs of maintaining those workers falls on the general taxpayers.

Governments are routinely involved with efforts to make certain energy-related activities safer for workers. This is most prominent regarding coal and nuclear fuel cycles, where dedicated government agencies exist to inspect and educate mines and production sites. If these costs are not paid entirely by the producers or consumers of the affected energy type, subsidies ensue. Public responsibility for workers' health care and/or pension costs also constitute subsidies. This has been quite common in the area of coal. For example, government payments to U.S. coal miners afflicted with black lung have exceeded \$30 billion. Black lung levels are now rising (or are being better documented) in other countries such as Russia, Ukraine, and China. Coal mine fatalities continue at extremely high levels in many of these countries as well.

Although difficult to quantify and normalize, external costs arise in many different ways through the fuel cycle. Table V provides an overview of these costs, in large part relying on data developed by ExternE. A project of the European Commission, ExternE has more than 50 research teams trying to quantify the externalities for key fuel cycles.

External costs are highest for coal and oil, although newer conversion technologies can reduce the toll per unit of power significantly. The environmental costs of nuclear power also appear low, even

TABLE V
Overview of Externalities by Energy Resource

Fuel	Years of life lost/TWh ^a	Environmental damage costs (Euro-cents/kWh) ^a	Quantifiable external costs (Euro-cents/kWh) ^b
Coal, 1995	1065	13	2.6
Coal, >2000	113	4	3.4 (lignite)
Oil, 1995	830	10	
Oil, >2000	139	2.6	
Gas, 1995	161	2.4	1.1
Gas, >2000	29	1.7	
Nuclear <100years	1	0	0.2
Nuclear >100years	8	0.2	
Biomass, range	20–100	0–0.8	
Photovoltaic, range	4–11	0.1–0.3	0.8
Wind	4–9	0	0.09
Hydro			0.07

^aRabl, A., and Spadaro, J. (2001). The ExternE project: Methodology, objectives, and limitations. In “Externalities and Energy Policy: The Life Cycle Approach”, Nuclear Energy Agency, workshop proceedings. Organization for Economic Cooperation and Development, Paris.

^bIncludes available estimates for health effects crop losses, material damage, noise, acidification/eutrophication, and global warming. Viridis, M. (2001). Energy policy and externalities: The life Cycle approach. Organization for Economic Cooperation and Development, Paris. In “Externalities and Energy Policy: The Life Approach”. Nuclear Energy Agency workshop proceedings.

though they are quite sensitive to assumptions regarding accidents at plants. The data also indicate that renewable energy sources would become far more competitive if the market were forced to account for external costs.

4. PATTERNS OF SUBSIDIZATION

Developing an aggregate picture of energy subsidization is extremely difficult due to the scope of policies affecting the sector and to tremendous fragmentation of the data across thousands of government ministries worldwide. There has been no global multifuel estimate of energy subsidy magnitude, although two efforts come close. Myers and Kent developed a global estimate for fossil (\$119 billion/year) and nuclear energy (\$12 billion/year), with an additional \$200 billion per year in externalities. Van Beers and de Moor estimated global subsidies to all fuels (\$250 billion/year, with more than 60% supporting fossil) but did not include externalities. Both estimates were built by aggregating data from preexisting smaller assessments rather than from a comprehensive policy survey. Different definitional and valuation approaches, as well as gaps in regions covered, make aggregation difficult. Although the resultant values indicate that energy subsidies are quite large, they are

unlikely to provide much precision as to the true subsidy magnitude.

Common themes are nonetheless visible across many of the smaller studies. For example, Koplow and Dernbach found that the fossil share of total subsidies exceeded 55% in four of six major U.S. multifuel energy subsidy studies conducted since 1978. In the remaining two, methodological inconsistencies largely explain the deviation. Within the fossil category, coal and/or oil normally received a larger share than did cleaner natural gas. Estimates for nuclear varied more widely. Where liability caps for accidents and capital subsidies were included, nuclear subsidies comprised 25% or more of the totals. Where they were incorrectly excluded, reported nuclear subsidies were much lower. Renewables and efficiency garnered 5% or less in most of the studies. Only when many large subsidies to other fuels were excluded did the aggregate subsidy estimates fall low enough to bring the renewables/efficiency share above 15%.

This pattern across fuels is broadly consistent with trends discussed earlier in this article. Historical spending on R&D has heavily favored fossil fuel and nuclear power, as has the use of export credit agencies. Subsidies to capital formation have supported primarily these same fuels, whereas those to energy security have favored oil and nuclear energy.

Failure to internalize pollution or health externalities generates subsidies primarily to coal and oil. Yet for many large industrial consumers, these fuels are exempt from taxation throughout the OECD. Government support for renewables, efficiency, and conservation is extremely modest. There have been recent gains in the share of public R&D supporting these areas, and some newer energy resources are eligible for tax credits or are included within purchase mandates. However, the overall pattern is one in which established energy types, often with poor environmental profiles, continue to receive most of the public support provided to the energy sector.

5. SUBSIDY REFORM

Modeling efforts, summarized by Koplow and Dernbach, suggest that subsidy reform is likely to generate environmental gains. Within the United States, modeling suggests that reforms could achieve a substantial (18% of the nonsequestration reductions) portion of the Kyoto 2010 carbon reduction targets. Modeling of international reforms projects some global reductions in greenhouse gas emissions by 2010 (0.2–8.0%), although results are sensitive to assumptions regarding subsidy values and fuel substitution options. Given the hundreds of billions of dollars involved with current subsidies, reforms also seem likely to reduce fiscal expenditures in many countries throughout the world.

Energy subsidies and subsidy reform have attracted the attention of many international governmental and nongovernmental organizations, as well as country-level ministries, over the past 15 years or so. Despite this attention, successful reforms have been few. This outcome reflects the role of political economy in subsidy creation, continuation, and reform. Who gets public resources is often an intensely political decision. Powerful groups in society are best positioned to institute policies that generate transfers to themselves. Furthermore, the theory of rent seeking indicates that groups that have received subsidies in the past will invest at least a portion of those gains to ensure that the subsidies keep coming.

Fossil, nuclear, and hydro energy sources all have been around for a long time, involve large companies and/or large government ministries, and have sufficient scale to dedicate staff to political lobbying. In addition, because the cost of many subsidies rises as the installed base eligible for them grows, the large installed base of fossil, nuclear, and hydro also

contributes to these resources capturing the lion's share of subsidies.

Given the strong political opposition to subsidy reform, a transitional process to precede policy change with much increased transparency makes sense. Initial steps to qualitatively identify and describe subsidies seems simple but can greatly change the political dynamics of subsidization by making recipients more visible to their competitors and the taxpayers. Quantifying the value of these transfers is the next step and helps policymakers to prioritize which subsidies are most important to address first. Consistent approaches should be used so that estimates done for different countries, or those done by different researchers, can be compared more easily. Separating externalities from fiscal subsidies is one element of standardization, as is applying consistent rules on offsets.

Modeling the impacts of these transfers on human, environmental, and fiscal health should come next, followed by modeling the effects of potential subsidy reforms. Summarizing a series of global workshops on energy subsidies, the UN Environment Program suggested that reforms should be implemented gradually and with strong transitional policies. A gradual approach might not make sense in all cases given that political forces may be able to stop some reforms entirely. However, transitional approaches that decouple payments from practices that harm human health or the environment should be pursued. Eligibility for existing subsidies can also be made contingent on acceptable environmental practices. Finally, new or replacement subsidies should be structured to leverage competitive markets (as do the RPSs) rather than providing support whether or not there is a successful outcome.

6. CONCLUSION

Subsidies remain a large presence in energy markets throughout the world. The scope, complexity, and politics of these policies help to explain why there is no global subsidy data set. However, all indications are that these subsidies cost hundreds of billions of dollars per year, impede market penetration of cleaner and more efficient methods of providing energy services, and increase damages to human health and the environment. Efforts to overcome the inherent political resistance to subsidy reform are needed, if only to greatly improve the ability to identify, describe, and quantify subsidies to particular fuels throughout the world.

Acknowledgments

Section II.C is from Koplow and Dernbach (2001) and is reprinted with permission from the *Annual Review of Energy and the Environment*, volume 26. Copyright 2001, Annual Reviews.

SEE ALSO THE FOLLOWING ARTICLES

Corporate Environmental Strategy • Market-Based Instruments, Overview • National Energy Policy: United States • National Security and Energy • Research and Development Trends for Energy • Strategic Petroleum Reserves

Further Reading

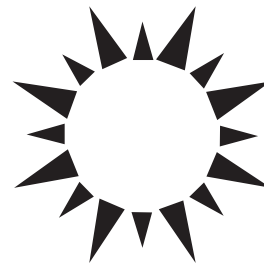
- Anderson, E., and Talley, W. (1995). The oil spill size of tanker barge accidents: Determinants and policy implications. *Land Econ.* 71, 216–228.
- Cone, B., et al. (1978). “An Analysis of Federal Incentives Used to Stimulate Energy Production.” Rep. PNL-2410REV. Pacific Northwest Laboratory, Richland, WA.
- Heyes, A. (2002–2003, Winter). Determining the price of Price–Anderson. *Regulation*, pp. 26–30.
- International Energy Agency. (1999). “World Energy Outlook: Looking at Energy Subsidies—Getting the Prices Right.” IEA, Paris.
- International Energy Agency. (2001). “Nuclear Power in the OECD.” IEA, Paris.
- Koplow, D. (1993). “Federal Energy Subsidies: Energy, Environmental, and Fiscal Impacts, main report and technical appendix.” Alliance to Save Energy, Washington, DC.
- Koplow, D. (1996). “Energy subsidies and the environment. In Subsidies and Environment: Exploring the linkages.” Organization for Economic Cooperation and Development, Paris.
- Koplow, D., and Dernbach, J. (2001). Federal fossil fuel subsidies and greenhouse gas emissions a case study of increasing transparency for fiscal policy. *Annu. Rev. Energy Environ.* 26, 361–389.
- Koplow, D., and Martin, A. (1998). “Fueling Global Warming: federal subsidies to oil in the United States.” Greenpeace, Washington, DC.
- Litman, T. (2002). “Evaluating Transportation Land Use Impacts.” Victoria Transport Policy Institute, Victoria, British Columbia.
- Lovins, A., and Lovins, H. (1982). “Brittle Power.” Brickhouse Publishing, Andover, MA.
- Margolis, R., and Kammen, D. (1999). Evidence of underinvestment in energy R&D in the United States and the impact of federal policy. *Energy Policy* 27, 575–584.
- Myers, N., and Kent, J. (2001). “Perverse Subsidies: How Tax Dollars Can Undercut the Environment and the Economy.” Island Press, London.
- Organization for Economic Cooperation and Development. (1998). “Improving the Environment through Reducing Subsidies,” part II: “Analysis and Overview of Studies.” OECD, Paris.
- Organization for Economic Cooperation and Development. (2001). “Environmentally Related Taxes in OECD Countries: Issues and Strategies.” OECD, Paris.
- Rabl, A., and Spadaro, J. (2001). The ExternE project: Methodology, objectives, and limitations. In “Externalities and Energy Policy: The Life Cycle Analysis Approach,” Nuclear Energy Agency workshop proceedings. Organization for Economic Cooperation and Development, Paris.
- UN Environment Program and International Energy Agency. (2001). “Energy Subsidy Reform and Sustainable Development: Challenges for Policymakers,” synthesis report, submission to the 9th session of the UN Commission on Sustainable Development.
- van Beers, C., and Moor, A. (2001). “Public Subsidies and Policy Failures: How Subsidies Distort the Natural Environment, Equity, and Trade and How to Reform Them.” Edward Elgar, Northampton, MA.

Suburbanization and Energy

DERON LOVAAS

Natural Resources Defense Council

Washington, D.C., United States



-
1. Suburban Sprawl
 2. Transportation in Suburbia
 3. Structures in Suburbia
 4. Land Use, Transportation, and Energy Consumption in Other Countries
 5. Transportation Alternatives for U.S. Suburbs

Glossary

developed land As defined by the U.S. Department of Agriculture, developed land is any land no longer a part of the rural land base of the United States. This includes large tracts of urban and built-up land, tracts of built-up land of less than 10 acres, and land outside of built-up areas that is part of a transportation system (i.e., roads, rail, and associated rights-of-way).

metropolitan statistical areas As defined by the U.S. Office of Management and Budget and used by the U.S. Census, this is the delineation of urban areas that includes the central city and any surrounding counties that are socially and economically connected to an urban center.

smart growth Alternative forms of development that use land efficiently, mix land uses (such as housing, business, and shopping), and encourage walking and public transit in addition to driving while protecting surrounding open space.

sprawl Low-density, automobile-dependent development with a strict separation of land uses.

vehicle miles traveled Surface transportation miles traveled annually by passenger and freight vehicles in the United States.

Shifting from the 19th to the 20th century, new forms of transportation reshaped U.S. urban areas. Streetcars and subway lines transformed cities such as Boston, New York, and Chicago, allowing for dense concentrations of businesses and hotels. These areas were surrounded by neighborhoods often bound by ethnicity and class.

Idealists planned expansions of the city as a whole through movements such as the City Beautiful movement and the Garden Cities movement as well as through model new suburbs such as Radburn, New Jersey, Mariemont, Ohio, and Greenbelt, Maryland. These designs were precursors for those built throughout the country to fulfill the American Dream of owning a single-family detached home and an automobile. The dream became reality for millions of Americans in the years after World War II, and construction accelerated due to incentives such as the home mortgage deduction and infrastructure investments.

Most notable in terms of its impacts on our energy consumption have been investments in highways needed to serve the form of transportation that would eclipse transit—the automobile. Additionally, the large structures connected by ribbons of asphalt in suburbia have consumed increasingly more energy. This article discusses suburbia's ascendance in the United States, the transportation and construction impacts of this trend, as well as some alternatives to sprawl being debated including “smart growth.” Smart growth entails new policies and incentives to create a greater variety in the supply of modes of transportation coupled with more and better travel demand management to conserve energy and improve environmental quality.

1. SUBURBAN SPRAWL

Much of current land development takes a form called suburban sprawl. Sprawl can be roughly defined as low-density development with a strict separation of land uses connected mostly (or only) by roads and highways. Although some sprawl occurred in the form of streetcar suburbs in the first couple of decades of the 20th century, sprawl accelerated after World War II as returning G.I.s, their families, and

TABLE I

Census Population Data for the 12 Largest Cities in the United States, 1950–1970^a

City	1950		1960		1970	
	Population	Rank	Population	Rank	Population	Rank
New York	7,891,957	1	7,781,984	1	7,894,862	1
Chicago	3,620,962	2	3,550,404	2	3,366,957	2
Philadelphia	2,071,605	3	2,002,512	4	1,948,609	4
Los Angeles	1,970,358	4	2,479,015	3	2,816,061	3
Detroit	1,849,568	5	1,670,144	5	1,511,482	5
Baltimore	949,708	6	939,024	6	905,759	7
Cleveland	914,808	7	876,050	8	750,903	10
Saint Louis	856,796	8	750,026	10	622,236	18
Washington, DC	802,178	9	763,956	9	756,510	9
Boston	801,444	10	697,197	13	641,071	16
San Francisco	775,357	11	740,316	12	715,674	13
Pittsburgh	676,806	12	604,332	16	520,117	24

^aData from Teaford (1993).

other citizens left cities for the suburbs in increasing numbers. This led to declines in the populations of central cities during the latter half of the 20th century (Table I).

A study by Janet Rothenberg Pack for the Brookings Institution examined population trends in 277 metropolitan statistical areas and found that flight from existing urban areas occurred throughout the United States from 1960 to 1990. The findings are summarized in Table II. In addition to the city-to-suburb migration evidenced by Table II, it is also clear that there have been interregional population distribution changes. Southern and western regions experienced the most population growth during these three decades. The population of suburban areas of these regions more than doubled from 1960 to 1990 (Table II).

Evidence shows that these population trends are continuing and that land development is occurring at twice the rate of population growth. Some of the best U.S. land-use data collection is done by the U.S. Department of Agriculture under the auspices of the Natural Resources Inventory (NRI). The NRI defines developed land as land that is no longer a part of the rural land base of the United States. This includes large tracts of urban and built-up land, tracts of built-up land of less than 10 acres, and land outside of built-up areas that is part of a transportation system (i.e., roads, rail, and associated rights-of-way). This land specifically consists of cropland, pasture land, range land, forest land, land enrolled in the federal Conservation Reserve Program, and other

TABLE II

Population Data (in Thousands) for U.S. Metropolitan Statistical Areas (MSAs), Cities and Suburbs, Divided by Region, 1960–1990

Area	Total population				% change, 1960–1990
	1960	1970	1980	1990	
All MSAs	127,575	149,125	164,042	182,512	43
Cities	56,529	60,568	59,279	62,114	10
Suburbs	71,046	88,557	104,764	120,398	69
Northeast	39,173	43,229	42,752	44,116	13
Cities	16,142	15,705	13,896	13,892	–14
Suburbs	23,031	27,524	28,856	30,224	31
Midwest	36,264	41,004	42,094	43,149	19
Cities	17,227	17,516	15,787	15,284	–11
Suburbs	19,036	23,489	26,307	27,865	46
South	30,625	37,278	45,463	53,349	74
Cities	14,207	16,695	17,551	18,506	30
Suburbs	16,418	20,583	27,912	34,843	112
West	21,513	27,613	33,733	41,898	95
Cities	8,952	10,653	12,045	14,431	61
Suburbs	12,561	16,961	21,689	27,467	119

rural land. According to the NRI, U.S. land development has accelerated from 1.4 million acres per year in the 1980s (1982–1992) to 2.1 million per year in the 1990s (1992–1997).

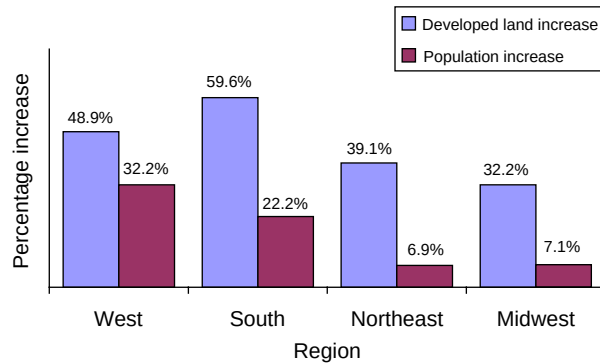


FIGURE 1 Population vs urbanized land growth by region, 1982–1997, as summarized in “Who Sprawl’s Most? How Growth Patterns Differ across the U.S.,” prepared in July 2001 by W. Fulton *et al.* for the Brookings Institution Center for Urban and Metropolitan Policy.

There are important regional differences in the relationship between these two factors in the sprawl equation. In the past couple of decades, the west has been growing but not sprawling as much as it has in the past (due in part to limits of developable land, whether due to topography or federal ownership), the northeast and the midwest have been sprawling but not growing much, and the south continues to grow and sprawl rapidly (Fig. 1).

2. TRANSPORTATION IN SUBURBIA

According to the U.S. Department of Transportation (DOT), as of 1995, every U.S. driver spent an average of 443 h every year behind the steering wheel, or 43 h more than in 1990. Although commuting to and from work only accounts for one-fifth of the miles that Americans drive daily, it is also notable that peak commuting hours have been increasing. In 1982, the average daily congested period was 2 or 3 h, but by 1999 this had doubled to 5 or 6 h.

There are economic impacts as well. Americans now spend more on transportation than food, clothing, or health care. As of 1998, according to the U.S. Department of Labor, 17.9¢ from every household dollar is spent on transportation, 98% of which is spent on buying, operating, and maintaining cars. This household expenditure is second only to housing (19¢). Transportation is also sapping the strength of the American economy. The Texas Transportation Institute found that congestion cost 75 metropolitan areas \$67.5 billion due to lost income and increased truck operating costs as well as excess fuel consumption (5.7 billion gallons) in 2000.

During the latter half of the 20th century, U.S. planners, engineers, and decision makers relied mostly on the expansion of road system capacity to alleviate traffic congestion. As quoted in “Divided Highways,” a history of the U.S. interstate highway system, in June 1969 Transportation Secretary John Volpe commented on this problematic modus operandi:

The federal government spends as much money on highway construction in 6 weeks as it has put into urban transit in the last 6 years. ... Unless we intend to pave the entire surface of the country—and no one wants that—we have to stop this trend. We already have 1 mile of highway for every square mile of land area in the USA.

According to the DOT, 30 years later the United States had 2 miles of highway for every square mile of territory. This country has made massive investments in building a 4-million-mile highway and road network since World War II. According to a 1997 National Research Council report, U.S. highways, streets, and associated rights-of-way totaled approximately 20 million acres, or 1% of U.S. land area. Based on this figure, the Harvard School of Design estimated roads impact one-fifth of the U.S. land base in the forms of fringe noise, air, and water pollution effects from vehicle traffic.

The environmental impacts of our growing road system are significant. Habitat loss and fragmentation reduce wildlife population diversity and abundance, contributing to species imperilment. In fact, the National Wildlife Federation reported that sprawl is the main cause of species imperilment in California. Wetlands loss can also be traced to our expanding transportation system: in the Environmental Protection Agency’s (EPA) “National Water Quality Inventory: 1996 Report to Congress,” 6 states reported significant wetlands loss due to highway construction and 10 reported significant loss due to residential development. This report also found that 38% of assessed U.S. estuary miles are impaired and that 46% of this impairment is due to polluted urban runoff. This is not surprising. Runoff from impervious surfaces such as roads is considerable: Runoff from a 1-acre parking lot is 16 times that of an undeveloped meadow.

Americans have made good use of increases in pavement. U.S. vehicle miles traveled (VMT) increased steadily throughout the second half of the 20th century. According to the DOT, from 1970 to 1990 VMT doubled from 1 to 2 trillion per year, and as of 1998 Americans traveled 2.6 trillion miles annually.

Driving increasingly more miles has meant turning away from other modes of transportation, such as mass transit. Using data originally from the DOT and the American Public Transportation Association and subsequently provided by the Surface Transportation Policy Project, Fig. 2 highlights this relationship for most of the 20th century. Transit boardings hit their peak in 1946 at approximately 23.5 million, whereas annual VMT decreased during World War II (1943–1945) and then continued its climb. This brief reversal of centurywide trends can be attributed to wartime policies such as the rationing of gasoline. In the 1920s, the larger trend lines had already taken shape. At century's end, however, VMT was barely growing, whereas transit stabilized and grew steadily from 1996 to 2001. The latest data available from 2002 show a decline in transit use of 1.3% and an increase in VMT of 1.7%. One overall trend is clear: Americans are driving increasingly more every year (Fig. 2).

2.1 Transportation Impacts on Energy and Air Quality

All of this driving has a major impact on U.S. petroleum consumption. According to the DOT, the transportation sector consumes more than 65% of the petroleum used in the United States and highway vehicles account for approximately 84% of this consumption. Basically, Americans are consuming increasingly more oil just to get around, despite improvements in efficiency. Stagnating fuel economy of the U.S. vehicle fleet due to the popularity of sport utility vehicles and light trucks doubtless contributes

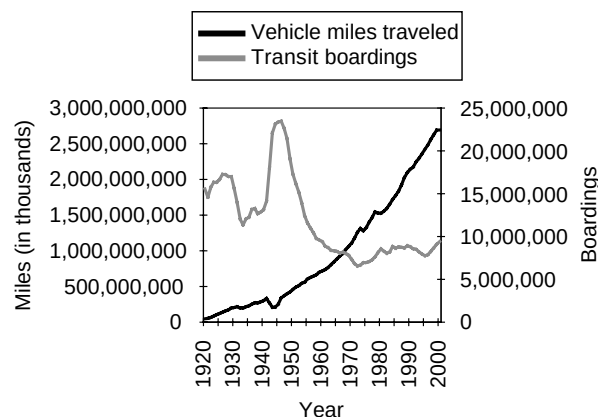


FIGURE 2 Change in transit boardings and vehicle miles traveled between 1920 and 2000. Data originally from the U.S. Department of Transportation and the American Public Transit Association as provided by the Surface Transportation Policy Project.

to this trend. According to the EPA, U.S. fuel economy peaked in 1987–1988 and is currently 6% lower than in 1988. Using U.S. Department of Energy (DOE) data, Fig. 3 shows that oil consumption by the U.S. transportation sector closely follows VMT growth—steadily upward toward the 9 million barrels per day mark.

Unfortunately, combustion of all this oil also pumps many unhealthy pollutants into the air. In addition to carbon monoxide (CO), cars and trucks are also the source of volatile organic compounds (VOCs) and oxides of nitrogen (NO_x). When VOCs and NO_x mix while exposed to the hot sun, a chemical reaction takes place producing ozone. In recent years, ground-level ozone has been particularly problematic for large, sprawling U.S. cities such as Atlanta and Houston. Ozone is an irritant to the lungs and eyes, causing coughing and chest pains. It aggravates asthma and can have especially harmful effects on children, who breathe more air per pound of body weight than adults.

Transportation is responsible for many of these emissions, as Benfield *et al.* found in their 1999 book, “Once There Were Greenfields”:

In particular, cars and other highway vehicles continue to emit some 60 million tons of carbon monoxide per year, about 62% of our national inventory of that pollutant; cars and other highway vehicles continue to emit some 7 million tons per year, almost 26%, of our volatile organic compounds ... and they emit around 8 million tons per year, about 32%, of our nitrogen oxides.

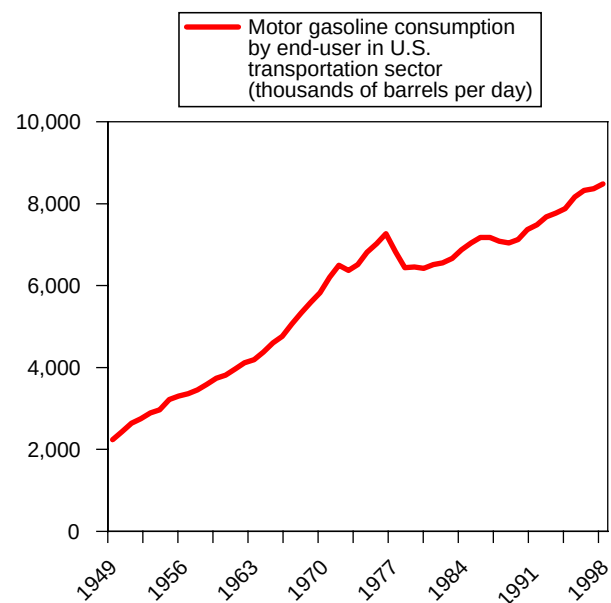


FIGURE 3 Increase in oil consumption by U.S. transportation sector, 1949–2001. Data from the U.S. Department of Energy.

The transportation sector is also a growing part of the total anthropogenic greenhouse gas emissions, especially carbon dioxide. According to the EPA, transportation sources accounted for approximately 31% of U.S. carbon emissions (460.4 million metric tons of carbon) in 1997. This sector is growing rapidly: Total emissions increased 10% between 1990 and 1997. According to the DOE, more than half of this increase can be traced to cars and trucks: As of 1999, 60% of transportation-related emissions were due to the combustion of gasoline.

3. STRUCTURES IN SUBURBIA

Although the impact of a transportation sector that underwent near-complete transformation in the 20th century has been examined most closely by experts and scholars, the low-density patterns of development have also meant an increase in the average size of houses, schools, and commercial buildings. This too has energy impacts.

3.1 Residential Buildings

According to the DOE, housing is the second largest source of carbon emissions in the United States. In 1999, residential buildings accounted for 19% of total U.S. CO₂ emissions at 290.3 million tons of carbon equivalent. According to the DOE, electricity accounted for most (63%) household carbon emissions as of 1997. Not surprisingly, the large-lot, detached-housing-only residential units that make up an increasing part of the U.S. landscape consume a great deal of electricity (as well as other types of energy) compared to more modestly sized housing types. In fact, according to the DOE, as of 1997 single-family detached housing made up by far the largest percentage of residential sector energy consumption at 73.4%.

Table III summarizes consumption levels for different U.S. housing types in 1997.

A good measurement for the increase in housing size in the United States is the number of rooms in a unit. The DOE's 1997 Residential Energy Consumption Survey showed that the size of units has grown steadily larger, resulting in greater energy use due to increased requirements for heating, air-conditioning, as well as lighting. Figure 4 shows this trend from 1978 to 1997.

Another yardstick for comparing the environmental impacts of different housing types is the ecological footprint of a given residence. Ecological footprint analysis is a tool designed by William Rees of the University of British Columbia and Mathis Wackernagel of the nonprofit Redefining Progress in the 1990s. This framework basically measures the depletion of "natural capital" by calculating how much world-average bioproductive land is needed to sustain a given population. The tool is versatile and has been used by households, schools, cities, and even nations.

This method of analyzing the impact of human activity sometimes uses the loss of "sinks" (soil,

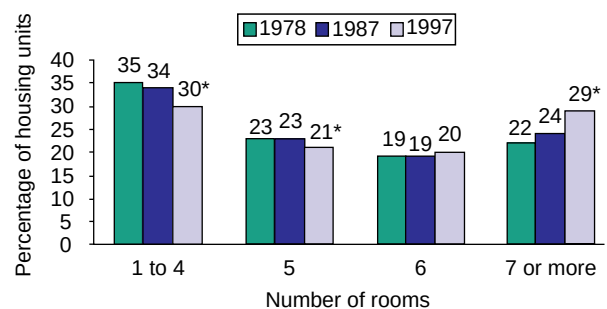


FIGURE 4 Increases in the number of rooms per household, 1978–1997. *The difference between the 1978 and 1997 estimates is statistically significant at the 95% confidence level. Data from the Energy Information Administration, Residential Energy Consumption Survey, 1978, 1987, and 1997.

TABLE III

Energy Consumption by Household Type in the United States, 1997^a

	Type of housing unit				
	All households	Single-family, detached	Single-family, attached	Multifamily, 2–4 units	Multifamily, 5 + units
Electricity consumption per household (kWh)	10,219	11,778	8071	6505	5990

^aData from the Energy Information Administration.

vegetation, and water resources that absorb CO₂, removing it from the atmosphere) to account for fossil fuel consumption. It also includes the bioproductive-land-equivalent impacts of food, the built environment, and forest. Overall, it is a useful sketch tool for comparing impacts, and an analysis of housing footprints yields interesting results. Specifically, Table IV provides the results of a 1995 study of different Canadian housing types using this tool.

Suburban sprawl includes a large number of single-family detached houses. For every detached house built instead of a townhouse, ecological footprint per capita increases 17%. Also, for every detached house built instead of an apartment, a person's ecological footprint increases 40%.

As the awareness of the significant impacts of residential development on the environment has grown, a cottage "green building" industry has also grown. Several U.S. jurisdictions and groups have codes, departments, or certification programs for green building. Austin (Texas), Boulder (Colorado), San Francisco, Seattle, Dade County (Florida), Oregon, Pennsylvania, New Jersey, New York, Minnesota, Maryland, Colorado, and Texas all feature green building efforts of some kind. Among private sector groups, the National Association of Homebuilders, the American Institute of Architects, the American Lung Association, and the Energy Efficient Building Association have programs. The U.S. Green Building Council is also working with nongovernmental organizations such as the Natural Resources Defense Council (NRDC) to develop standards for residential development as part of its Leadership in Energy and Environmental Design (LEED) certification process.

This increased interest in green building is driven in part by increasing demand. For example, 89% of U.S. consumers surveyed in a 2000 *Professional Builder* magazine poll said they "were willing to pay more for green building features if they improved quality, durability, and the health of the home."

More than half would pay an additional \$2500–5000 and 9.5% would pay an additional \$10,000 for a green home.

Considerable energy savings are possible through efficient design. In a 1998 study of the savings possible via cost-effective energy efficiency investments (in lighting, appliances, etc.) in Illinois, the American Council for an Energy Efficient Economy (ACEEE) found that 32 and 27.5% savings in electricity were achievable for existing and new single-family units, respectively. A similar 1997 ACEEE study found 37.5% (existing) and 26.5% (new) in possible savings in Pennsylvania and New York and a 51.5% (existing) and 30.0% (new) possible savings in New Jersey, all for single-family units. Therefore, possible electricity savings through energy efficiency improvements for the type of housing typically found in sprawling developments in these four states ranges from 26.5 to 51.5%.

However, few green building programs integrate criteria about neighborhood density and location efficiency, which are key components of smart growth. Changes are afoot, though. For example, Austin, Texas, has a Smart Growth Initiative complete with a matrix for calculating just how "smart" particular development proposals are, and LEED is developing new design standards that integrate "smart growth" criteria. Given the popularity of the approximately 100 New Urbanist developments built thus far, the addition of location efficiency as a criterion for determining how "green" a building is should ensure that such buildings continue to fetch a premium in the marketplace.

3.2 Educational Buildings

Buildings in the U.S. educational system comprise approximately 13% of total commercial building floorspace. According to the DOE, they are the largest at 25,100 square feet on average per building, compared to the average for all commercial buildings

TABLE IV
Footprints of Various Household Types in Acres^a

	Category				
	Standard lot detached house	Small lot detached house	Townhouse	Walk-up apartment	High-rise apartment
Housing	1.95	1.95	1.61	1.16	1.16

^a Source: Walker and Rees (1997). For more information, see <http://www.ecologicalfootprint.org>.

of 12,800 square feet. Newer schools use more energy per square foot than older schools, according to the Energy Information Administration. The National Trust for Historic Preservation finds that newer schools are larger and more apt to be located in sprawl locations than older ones. In 2000, the Department of Education studied the functional age of schools, or the number of years since a building had been constructed or renovated, and found that newer schools were larger:

On average, small schools had older functional ages than medium or large schools. Additionally, large schools were more likely than small schools to have a functional age of less than 5 years, and the small schools were more likely than large schools to have a functional age between 15 and 34 years.

This is not surprising given requirements imposed on school construction by states and localities. For example, many states adhere to the guidelines created by the Council of Educational Facility Planners International (CEFPI) when planning new school construction:

- At least 10 acres of land plus 1 acre for every 1000 students for elementary schools
- At least 20 acres of land plus 1 acre for every 100 students for middle schools
- At least 30 acres of land plus 1 acre for every 100 students for high schools

Not surprisingly, buildings on this land are also larger than they have been in the past. In 1995, CEFPI released what it terms the “starting point” for planning the size of school buildings. Size of schools is partly determined by climatic conditions. For this reason, CEFPI lists different averages for currently existing school buildings in different regions. One region encompasses nearly the entire United States (48 states). For this region, the recommended starting point average gross square footages per student are 111.5 for elementary schools, 154.4 for middle schools, and 160.7 for high schools.

3.3 Commercial Buildings

According to the DOE, in 1999 commercial buildings consumed 16% of U.S. primary energy, and 75% of this consumption was in the form of electricity. As of 1995, the greatest energy consumers were buildings that provided or hosted food sales or services, health care-related services, lodging, public assembly, public order and safety services (i.e., police stations), and office spaces. According to the Alliance

to Save Energy, carbon dioxide emissions from the commercial buildings sector have increased 30% since 1980, faster than any other sector.

The DOE sums up the scale of our stock of commercial buildings as follows:

In 1995, there were 4.6 (with a ± 0.4 margin of error) million commercial buildings in the United States comprising 58.8 (± 3.9) billion square feet of floorspace. That amount of commercial floorspace exceeds the total area of the state of Delaware and amounts to more than 200 square feet for every resident in the United States.

In spite of the tremendous amount of land development in recent decades, this stock is surprisingly comprised mostly of older buildings. The median building age is 30.5 years, and more than half of the buildings were built before 1970. According to the DOE, site energy intensity has not decreased significantly since the beginning of the 20th century. A lack of progress in the 1990s is confirmed by a November 2000 assessment of energy consumption by the EPA, which states, “Total energy consumption in commercial buildings remained roughly constant from 1989 to 1995, as did energy use per square foot.”

Highlighting the potential to conserve energy, the EPA cites studies showing that “by investing in simple profitable building upgrades, many U.S. homes and businesses could reduce their annual energy use by nearly 30%.” Through both private and public sector programs, such savings have been achieved in some cases. The EPA’s Energy Star program, for example, claims to have partnered with 15% of the “commercial, public, and industrial building market” and rewarded 90 energy-saving buildings with the Energy Star label. These buildings prevent 45 million metric tons of carbon equivalent worth of emissions. Also, the Rocky Mountain Institute has performed many studies that show potential energy savings for commercial buildings: In one analysis, they found potential savings of 75% for a vacant 13-story, 200,200-square foot office building outside Chicago.

4. LAND USE, TRANSPORTATION, AND ENERGY CONSUMPTION IN OTHER COUNTRIES

Urban settlement patterns in other countries have remained fairly compact. Combined with adequate transportation infrastructure, this leads to lower dependence on the automobile in countries such as those in Europe. In fact, European land-use and

transportation policies and practices are often evoked when trying to define policy alternatives for achieving more effective growth management in the United States.

For example, Pietro Nivola of the Brookings Institution finds that U.S. cities interested in improving their condition could adopt some policies used frequently in Europe, such as reducing public funding for highway construction, reducing urban crime, and providing fiscal relief for city businesses and residents through revenue-sharing arrangements such as those used by Germany and Canada. Timothy Beatley of the University of Virginia has researched best practices in European countries, including Copenhagen's 1947 land-use "finger plan" that concentrates development into a hand shape, with the city center as the palm and transit-line-centered digits extending outward with protected green spaces between the lines. Due in part to the choices offered by these innovative land-use and transportation policies, 31% of Copenhagen commuters ride transit to work and 34% bike to work. This compares poorly to U.S. cities, where transit and bike home-to-work trips make up merely 5% and 2% of the total, respectively.

European cities are faring better in terms of fuel use. In 1999, Europe was second only to the United States (at 25%), with 15% of total global carbon emissions from fossil fuels. However, according to the DOE, European emissions per capita were less than half those of the United States (2.4 vs 5.6 tons).

A study of 48 global cities by Peter Newman and Jeffrey Kenworthy shows just how energy-intensive U.S. transportation patterns have become. The averages for gasoline used for private transportation are summarized in Fig. 5.

They found generally that gasoline-oriented cities are "heavy energy users while those with any

significant level of electricity use [to run a rail system] in their transportation system are low energy users overall." Even after correcting for economic factors, such as U.S. income, gasoline prices, and fuel efficiencies, they found that gasoline use was still significantly higher (47–63%) in the United States than in other world cities, on average. In fact, the authors noted that "a purely economic approach to transportation matters will be inadequate, and that matters of infrastructure and urban form have direct influence on transportation patterns." One advantage to investing in more transportation options in order to reduce trips and trip lengths is a phenomenon termed transit leverage. Newman and Kenworthy cite Neff, who found that transit does not simply replace car travel on a one-to-one ratio. In fact, Neff found that 8.6–12.0 km (5.3–7.4 miles) of car travel is replaced by 1 km (0.62 miles) of transit according to data from the United States.

Interestingly, Newman and Kenworthy also found that there is little connection between transportation and economic growth. Whereas in U.S. cities gross regional products per capita were only 0.85 those of Europe, car use was 2.41 times higher in the United States.

Nonetheless, the suburban sprawl model is being adopted by Second and Third World countries as a means to prosper. Private car use in Europe and China is increasing. As Molly O'Meara Sheehan of the Worldwatch Institute wrote,

In the Czech Republic, car use surged and public transit use fell as the number of suburban hypermarkets ballooned from 1 to 53 between 1997 and 2000. Today, the proliferation of such developments around Prague has boosted the number of hypermarkets in the Czech Republic above that in The Netherlands.

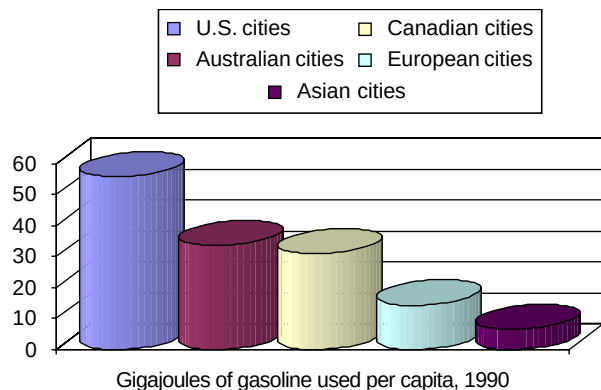


FIGURE 5 Gasoline use per capita in the United States, Canada, Europe, Australia, and Asia.

5. TRANSPORTATION ALTERNATIVES FOR U.S. SUBURBS

As discussed previously, since the mid-20th century popularization of the automobile, the usual response to demand for faster commutes and trips has been to provide a more ample supply of new highway and road capacity. However, in recent years various studies have described a phenomenon that puts in doubt the effectiveness of this tactic for reducing congestion. The phenomenon is not surprising: Greater availability of pavement spurs more and longer trips by road users. What is sobering is the extent of the phenomenon.

The influence of new road and lane construction on travel behavior is referred to as induced travel, which has been studied in all 50 states and 70 regions. The idea is simple: New road capacity acts as a magnet for drivers, inducing them to take more and longer trips than they would have otherwise. According to the growing academic literature as summarized in 2000 in a presentation by Lewison Lee Lem of the EPA, the impact can be substantial: An average increase of 10–30% of total VMT can be traced to induced travel. Short-term induced increases range from 2 to 5% for every 10% increase in road capacity. Long-term increases may result in increasing travel 5–10% for every 10% increase in capacity.

Some of this additional traffic is also a product of new land development and new users who occupy new subdivisions and shop in new malls. Transportation investments often have an impact on land development. A study performed for the Brookings Institution by Marlon Boarnett confirmed the relationship between transportation policy and land use, concluding that highways can induce changes in regional development patterns. In fact, this study found that highway investments shift economic development away from existing communities to newly built areas. Just as empty roads and traffic lanes act as magnets for new driving, highway interchanges draw new development.

Smart-growth policies and practices offer one possible way to deal with out-of-control VMT and pavement growth. These policies are being embraced by localities and states throughout the country. It is time for such a shift. Many surveys in the past few years have shown that Americans are tired of suburban sprawl and its symptoms—loss of open space, traffic congestion, and blight in older communities. For example, in a 1999 poll by the Pew Center for Civic Journalism, sprawl and its effects tied with crime as the issue of most importance at the community level.

What is smart growth? Roughly, smart growth is development that involves a mix of land uses (e.g., commercial and residential), which occurs in or adjacent to existing communities and which provides a variety of transportation choices for residents and workers. Orienting and designing new development to make it convenient for residents and workers to take transit is especially important, and there is even a name for this kind of development—transit-oriented development.

Many advocates propose taking advantage of significant infill development and redevelopment opportunities created by population declines in central cities and older suburbs and postwar leapfrog

development patterns. Retrofitting U.S. suburbs so that we take advantage of vacant and abandoned lots means more compactness and convenient access to shops and potential workplaces. After decades of sprawl, there are plenty of opportunities for such retrofits. One example that is receiving much attention is “greyfields.” These are the older regional malls that litter many older suburbs. Unfortunately, “litter” is an all-too-apt term since many of them are vacant, having been left behind by companies seeking to build ever-larger stores and malls. According to a study by PricewaterhouseCoopers and the Congress for New Urbanism, 140 of approximately 2000 regional malls are greyfields, and 200–250 are on their way to that status. These sites are perfect for redevelopment given their large size—35 acres on average—plus the fact that their location in older suburbs often means proximity to transit.

Redevelopment activities such as this will mean more jobs and more residents per acre, which many studies have found correlate well with decreases in VMT. A 1992 study by Cambridge Systematics found 20–25% fewer car trips per household in more compact, transit-oriented suburbs than in sprawling, auto-dependent suburbs. In a 1994 NRDC-sponsored study of several San Francisco communities, John Holtzclaw found that a doubling of residential density is accompanied by a 20–40% reduction in VMT. Also, Reid Ewing concluded in a 1997 study that when controlling for effects of variables such as income, a doubling of urban densities results in a 25–30% reduction in VMT.

It is important to clarify that when researchers say “doubling,” they are usually talking about modest regional or neighborhood averages for density or compactness. Some Americans might think that the only alternative to sprawl is New York City-level density. However, this is not the case. There are a number of alternatives to suburban sprawl, including some models within suburbia. For example, Ed Risse of Synergy Planning estimated in 2001 that if the Washington, DC region were to adopt the suburb of Reston’s density of 9 or 10 people per acre, the region would be able to accommodate 25 years of demand for urban development on vacant and underutilized land within a 20-mile radius of Washington’s center. Reston is only compact relative to other segments of suburbia: It remains a quiet, leafy community due in part to the 1300 acres of green spaces it has permanently set aside.

Some studies of the impacts of changes in the built environment on travel demand have yielded various results. In a 2001 synthesis of the literature, Reid Ewing and Robert Cervero concluded that travel

TABLE V

Summary of Environmental Protection Agency Findings Predicting Daily Vehicle Miles Traveled (VMT) and CO₂ Reductions in Infill/Redevelopment Building Scenarios

Location	Daily VMT (%). CO ₂ reduction (%)	
San Diego, CA	48 (per capita)	45
Montgomery County, MD	58 (per capita)	46
West Palm Beach, FL	61 (per capita)	50
Atlanta, GA	14.5–52	Not calculated

demand is pretty inelastic in the face of individual changes in the built environment, but that there are large cumulative effects, a combination of smart-growth characteristics, such as increases in density, mixing land uses, design changes, and wider regional accessibility of different land uses.

New, sophisticated modeling techniques have demonstrated the effectiveness of such strategies. Several EPA studies utilize four-step travel demand models to show that major reductions in VMT are achievable. These models use information about travel patterns to calculate how many trips will be generated, where trips will be distributed, and how the trips will be split between different transportation modes (i.e., walking, biking, train, and bus) and also to assign the traffic among different routes and times of day. In two studies, the EPA worked with local jurisdictions to choose likely sites for infill/redevelopment and greenfield development and used transportation models to forecast differences in miles traveled. Table V summarizes the findings.

Examining development patterns in the San Francisco, Los Angeles, and Chicago areas, NRDC researchers found strong correlations between density and urban form and VMT. In areas with smart-growth characteristics, such as small lot sizes, transit services, and walkable neighborhoods, residents find it less necessary to drive. This powerful relationship is expressed in Fig. 6.

Also, in EPA-commissioned studies of travel behavior in the Sacramento and Nashville regions, NRDC and Criterion Consultants found that VMT per capita in neighborhoods was lower in urban and older suburban neighborhoods. In Nashville, VMT per capita was 30% lower among suburbanites in an older, more compact suburb as opposed to a newer, more sprawling suburb. In Sacramento, the rate of fuel consumption was as much as 50% lower in an urban community compared to two suburban neighborhoods.

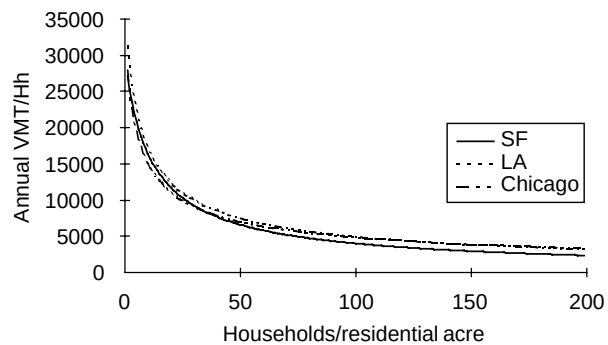


FIGURE 6 Relationship between VMT per household and residential density in the San Francisco (SF), Los Angeles (LA), and Chicago metropolitan areas.

It should be noted that there is great potential for synergy among several nascent market transformations, including smart growth. The combination of green building techniques, cleaner vehicle technologies such as hybrid engines and fuel cells, and changes in development of new suburbs and retrofits of existing communities should yield significant benefits in terms of energy conservation, climate change, and environmental quality.

5.1 The Return of Rail?

When we selected cars and trucks as our vehicles of choice, we shunned another kind of infrastructure—rail, especially intracity rail. Urban streets and even highways took the place of trolley systems that spread like webs through our cities. We had built approximately 45,000 miles of trolley lines by 1915. According to the American Public Transportation Association, we currently have a paltry 3100 miles of light and heavy rail serving our cities. As discussed previously, ridership on public transportation peaked at 23.4 billion trips in 1946 and declined steadily throughout the decades, bottoming out in 1972 at 6.5 billion.

However, there have been considerable increases in recent years and for 5 years in a row in the late 1990s the rate of transit boarding growth actually exceeded the growth rate of driving. The trend during the past 10 years is shown in Fig. 7.

In response to increasing demand, new rail starts have increased in recent years, including some starts in unexpected southern and western cities. As columnist Neal Peirce noted, applications for federal funding for new bus and rail projects leaped from 20 in 1997 to 191 in 1998. The federal program for new transit starts remains popular and hence oversubscribed. Also, the American Public Transportation Association reported that as of 2000, there

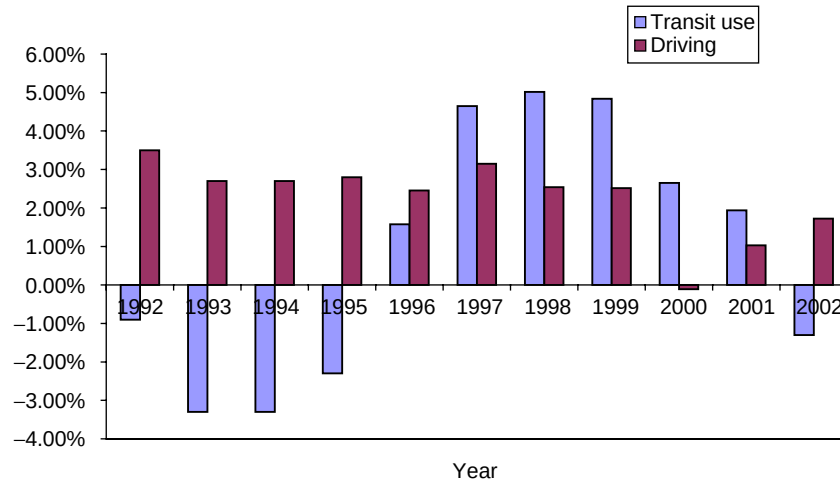


FIGURE 7 Percentage change in the growth of transit use and vehicle miles driven in the United States, 1992–2002. Data originally from the U.S. Department of Transportation and the American Public Transit Association as provided by the Surface Transportation Policy Project.

were approximately 1500 miles of new light rail and heavy rail lines in the proposal, planning, design, or construction phase.

This boom in new rail construction is a nationwide phenomenon. The following are examples from *Railway Age's* 2001 "Passenger Rail Planner's Guide":

- Charlotte, North Carolina, has completed a major investment study for an 11.5-mile light rail line that will share 2 miles of track with the town's trolley.
- Atlanta's Metropolitan Area Rapid Transit Authority (MARTA) is planning to extend its rail system with a 4.2-mile line, and work has begun on a 47-acre transit-oriented development on MARTA-owned land that will include BellSouth as a tenant.
- The Southwest Ohio Regional Transit Authority and the Transit Authority of Northern Kentucky are studying an 18-mile light rail segment that would connect Covington, Kentucky, and central Cincinnati.
- A 17.4-mile light rail extension in St. Louis was slated to open this year, and an additional 8.8-mile segment is planned.
- The Dallas Area Rapid Transit Authority is constructing two new lines for its successful system, which will more than double the system's current 20 miles.
- The Denver Regional Transportation District recently opened an 8.7-mile extension to its system and is planning another 19-mile line.
- Salt Lake City recently built a 15-mile light rail line that is wildly popular; several extensions are planned or under way.

Many jurisdictions devoting resources to building new transit are outside of America's Rust Belt and in the Sun Belt—the American south and west. Increasing rail investments can help to catalyze transit-oriented smart growth since concentrating development around rail nodes helps ensure their fiscal viability. In fact, the Federal Transit Administration includes land-use criteria when awarding Full Funding Grant Agreements for new transit service. New transit investments throughout the country provide evidence that smart-growth strategies are widely applicable—both for revitalizing northeastern and midwestern cities and older suburbs that have felt sprawl's impacts most keenly and for the rapidly growing southern and western United States.

In summary, the possible comeback of rail transportation in the United States is good news indeed for those concerned about 20th-century changes in development patterns that have transformed suburbanization, transportation, and energy consumption in the United States.

Acknowledgments

The author acknowledges Nancy Jakowitsch and Hannah Stutzman for their contributions to this article.

SEE ALSO THE FOLLOWING ARTICLES

Alternative Transportation Fuels: Contemporary Case Studies • *City Planning and Energy Use* •

Ecological Footprints and Energy • Heat Islands and Energy • Land Requirements of Energy Systems • Leisure, Energy Costs of • Lifestyles and Energy • Motor Vehicle Use, Social Costs of • Population Growth and Energy • Urbanization and Energy

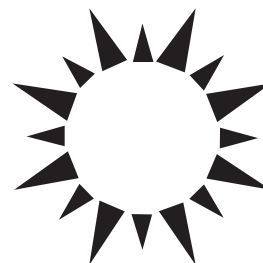
Further Reading

- Beatley, T. (2000). "Green Urbanism." Island Press, Washington, DC.
- Benfield, F. K., Raimi, M., and Chen, D. (1999). "Once There were Greenfields." Natural Resources Defense Council and Surface Transportation Policy Project, New York.
- Ewing, R., and Cervero, R. (2001). Travel and the built environment—A synthesis. *Transportation Research Record*.
- Forman, R. T. T., Sperling, D., *et al.* (2003). "Road Ecology, Science and Solutions." Island Press, Washington, DC.
- Gillham, O. (2002). "The Limitless City: A Primer on the Urban Sprawl Debate." Island Press, Washington, DC.
- Lewis, T. (1997). "Divided Highways." Penguin, New York.
- Newman, P., and Kenworthy, J. (1999). "Sustainability and Cities: Overcoming Automobile Independence." Island Press, Washington, DC.
- Nivola, P. (1999). "The Laws of the Landscape: How Policies Shape Cities in Europe and America." Brookings Institution Press, Washington, DC.
- O'Meara Sheehan, M. (2001). "City limits: Putting the brakes on sprawl, Worldwatch Paper 156." Worldwatch Institute, Washington, DC.
- Pack, J. R. (2002). "Growth and Convergence in Metropolitan America." : Brookings Institution Press, Washington, DC.
- Sierra Club (2000). "Sprawl Costs Us All: How Your Taxes Fuel Suburban Sprawl." Sierra Club, San Francisco.
- Teaford, J. C. (1993). "The Twentieth-Century American City." 2nd ed. Johns Hopkins Univ. Press, Baltimore.
- U.S. Environmental Protection Agency (2001). "Our built and natural environments: A technical review of the interactions between land use, transportation, and environmental quality," EPA 231-R-01-002. Development, Community, and Environment Division. U.S. Environmental Protection Agency, Washington, DC.
- Walker, L. A., and Rees, W. E. (1997). "Ecocity Dimensions: Healthy Communities, Healthy Planet" (M. Roseland, Ed.). New Society, Gabriola Island, British Columbia, Canada.

Sun, Energy from

WILLIAM LIVINGSTON

National Solar Observatory
Tucson, Arizona, United States



-
1. Solar Interior and Nuclear Fusion
 2. Magnetism and the Activity Cycle
 3. Total Solar Irradiance
 4. Solar Wind and Particle Flux
 5. Orbital Modulation
 6. Long-Term Variations in Solar Output

Glossary

astronomical unit (AU) The mean distance of the earth from the sun, which is 1.495979×10^8 km.

chromosphere Layer immediately above the photosphere where many strong Fraunhofer lines are in emission at the limb of the disk.

corona The extended upper atmosphere of the sun, normally seen only at the time of a total solar eclipse or from satellites with an occulting device to cover the much brighter white light solar disk; its outer bounds reach beyond the earth as the tenuous heliosphere.

corona hole Region on the solar disk where surface magnetic fields open out to allow particle streaming and subsequent magnetic storms on the earth.

faculae Bright markings on the solar disk especially evident toward the limb; faculae give rise to positive blips in the irradiance time record.

granulation The fine mottling pattern on the white light photosphere of the solar disk.

insolation Energy input to the earth's surface that depends on the geographic latitude, atmospheric absorption, and the like.

irradiance The amount of electromagnetic power that flows across a surface (W/m^2).

limb darkening The dimming of the solar disk toward its edge; it is an indicator of a photospheric temperature gradient that diminishes with height.

photosphere The solar disk in white light; most Fraunhofer lines are formed in the upper photosphere.

radiance The power radiated from a unit area on a planar source into a unit solid angle (W/steradian/m^2).

sunspots Dark markings on the solar disk, usually consisting of an umbra and the surrounding less dark penumbra.

transition layer A dynamic, high-temperature layer between the chromosphere and the corona; characterized by emission from ionized lines such as Si IV, C IV, and N V, it clearly displays the structure of magnetic field lines.

ultraviolet (UV) Radiation shorter than visible wavelengths between 0.4 and $0.2 \mu\text{m}$; extreme ultraviolet (EUV) is $< 0.2 \mu\text{m}$.

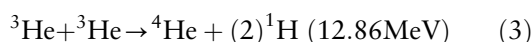
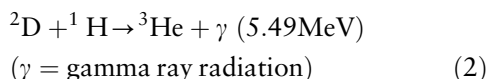
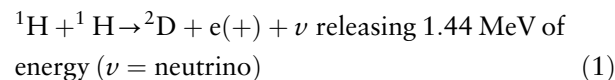
The sun is the primary source of energy on the earth. Most of this energy (99%) is in the form of electromagnetic radiation. Nuclear fusion in the sun's core is the origin of this energy. Space-borne radiometers have allowed the sun's radiation to be measured over the past quarter-century or so, and the result is constant to approximately 0.1%. This article considers these measurement techniques and their limitations. Ultraviolet emissions and particle flows, although relatively minute, are strongly modulated by the sun's 11-year magnetic cycle and produce noticeable effects on the earth's upper atmosphere. The long-term record of the sun's output and its magnetic cycle is indirectly revealed by historic sunspot records and from the geo-sited nuclides of ^{14}C and ^{10}Be . Other solar-type stars suggest other activity cycles that the sun might have exhibited in the past or might develop in the future.

1. SOLAR INTERIOR AND NUCLEAR FUSION

The chemical composition of the sun is 71% hydrogen, 27% helium, and 2% other elements. According to the Big Bang theory of the origin of the universe, only hydrogen and a few light elements were created. The presence of heavier elements, such as Fe, Ca, and Mg, indicates a stellar evolution that

includes material processing in the distant past through a supernova explosion.

Physical conditions in the solar interior are such that perfect gas laws apply. Therefore, one can calculate that at the sun's center the density is $\rho = 150 \text{ g/cm}$ and the temperature is $15 \times 10^6 \text{ K}$. At this density and temperature, the proton-proton (p-p) reaction takes place:



Thus, H is burned to produce He, and this is the source of the sun's luminosity ($L = 3.845 \times 10^{26} \text{ W}$). The age of the sun, which is approximately the age of the earth, is estimated to be 4.5×10^9 years. In the core, 50% of H has now burned.

Fusion energy generation is confined to the inner third of the solar radius, or $0.3 R_{\odot}$. Energy flows outward by radiation to $0.75 R_{\odot}$, where a convective zone of transport sets in via "giant cells." At the surface, the final convection elements are the white light granulation. Bright granules of observed mean dimension 1.4 seconds of arc (range 0.25–3.50, arcsec) are buoyed to the surface and radiate into space, and the resulting cooled gas (-300 K) descends as dark intergranule lanes (Fig. 1).

It takes approximately 10 years for core radiation (gamma rays or radiation whose wavelength is

$< 0.01 \text{ nm}$) to diffuse to the surface and degrade to a cooled gas (effective temperature 5777 K or visible radiation). Currently, core conditions can be sensed from the neutrino flux (in minutes), which verifies the forgoing model of the solar interior. Seismic waves, which pass through the sun in hours, are detected at the surface as small Doppler shifts of the Fraunhofer lines and likewise substantiate the solar model.

2. MAGNETISM AND THE ACTIVITY CYCLE

The sun is a largely passive body. In fact, the sun possesses magnetic fields whose surface manifestation waxes and wanes at approximately 11-year intervals. This introduces the dynamic aspects of the sun. Unlike the earth, with its buried magnetic dipole, the sun's magnetism is observationally a surface phenomenon. Subsurface magnetism must exist, but we know little about it directly. The sun does not rotate as a solid body but rather exhibits differential rotation, with a surface angular rate at the equator greater than that at higher latitudes. Theory proposes that this differential rotation acts as a dynamo to intensify any diffuse magnetism in the sun's convective zone. Once magnetism is concentrated by this dynamo, the associated gas is less dense than nonmagnetic surroundings and is buoyed to the surface, eventually emerging as dark sunspots. Sunspots are dark because convection is inhibited inside them. (It should be noted that subsurface gas is nearly completely ionized, a plasma that is a perfect electrical conductor, and that magnetic fields are "frozen in" and virtually indestructible. A fraction of the sun's magnetism may actually be a relic from when the sun condensed out of the solar nebula.)

Sunspots typically appear in bipolar pairs having field strengths in their dark umbrae of 2000 to 3500 Gauss. It takes only a few days for a large sunspot group to form. Because sunspots are dark, the sun's total radiation diminishes proportional to their area up to 0.1%. Afterward, surface motions "nibble" away at sunspots and carry off the fields by a random walk fashion to create faculae. Faculae have fields of 500 to 1000 Gauss and are relatively bright. Thus, dark sunspots transform to bright faculae that outlive the spots themselves. In this way, over time, sunspots enhance solar luminosity by approximately 0.1%.

Figure 2 is a "magnetogram" of the sun near cycle maximum. On this picture, light areas represent positive (north) fields, and dark areas represent

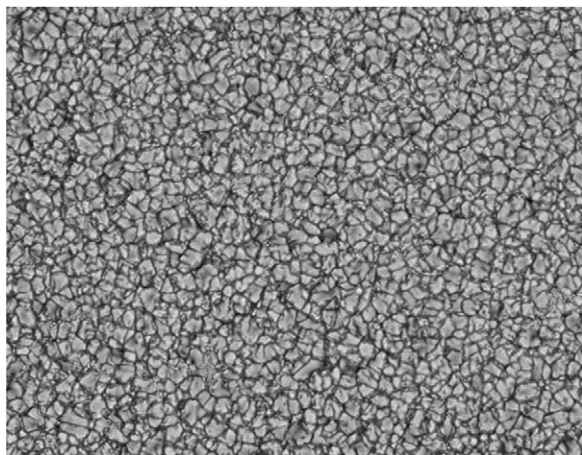


FIGURE 1 White light of solar granulation in a quiet (nonmagnetic) region. (Dutch Open Telescope, La Palma, Canary Islands, October 19, 2001)

negative (south) fields. The fragmented fields are cospatial with faculae. Most of the solidly light or dark areas are composed of unresolved faculae. Spots are not distinguishable here but lie within the shaded areas. Compare this figure with Fig. 3, which is the same type of picture at solar minimum. The fragmented fields in Fig. 3 are likewise faculae.

In 1843, Schwabe discovered that the number of spots on the solar disk is cyclic, with an average period of approximately 11 years. The sunspot cycle

is synonymous with the magnetic or activity cycle. Besides sunspot number, other “solar indexes” include the number of sunspot groups, 10-cm radio flux, and the strength of certain spectrum lines. Fraunhofer lines, such as NaD1 (596 nm) and CN (388 nm), report on the upper photosphere, CaK (393 nm) reports on the chromosphere, He (1083 nm) reports on the chromosphere and corona, and Lyman- α (121 nm) and other highly ionized lines, on the transition layer.

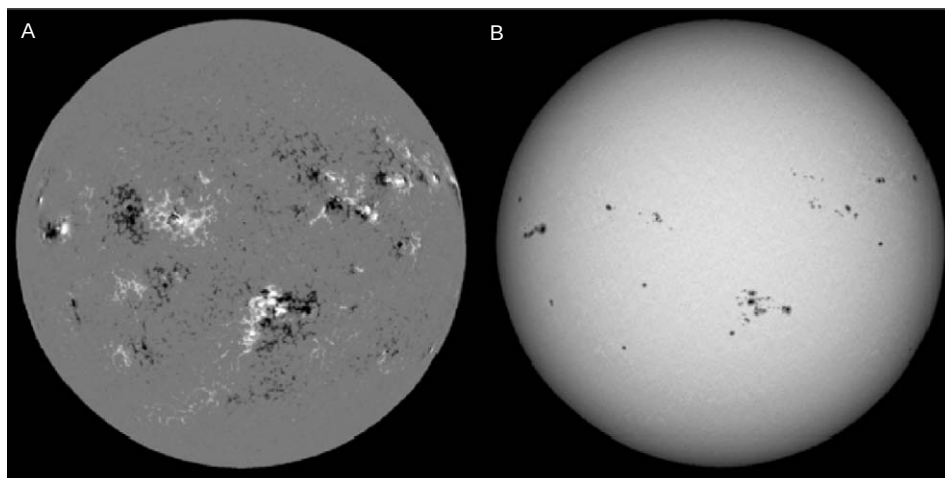


FIGURE 2 (A) A full-disk magnetogram taken near solar maximum on July 20, 2000. This is a computer-drawn picture showing polarization in the Fe (868.8-nm) line. When certain lines are formed in the presence of a magnetic field, Zeeman effect splitting occurs, with the amount being proportional to field strength, and the split components are polarized depending on magnetic polarity. A magnetograph measures these parameters. In the picture, white is north magnetic polarity, and black is south magnetic polarity. The east limb of the sun is on the left. The sun rotates east to west. The northern and southern activity belts are separated by the equator, which spans the middle. (B) A white light image for the same date.

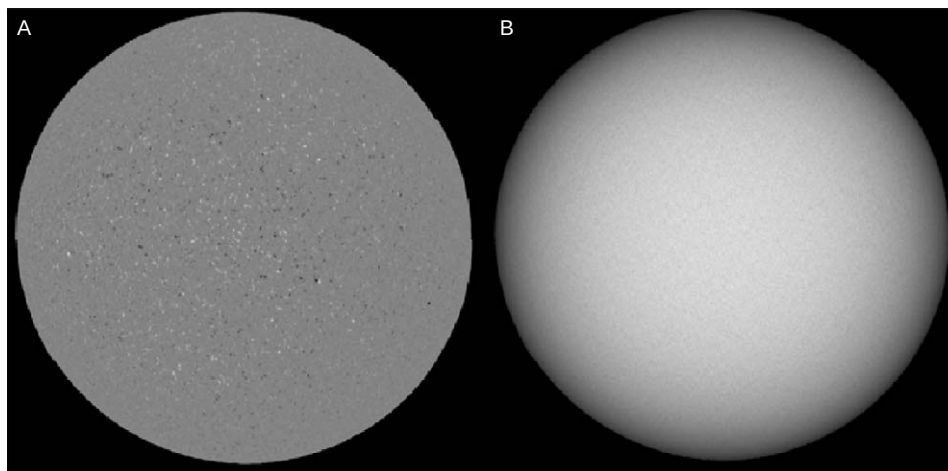


FIGURE 3 (A) A full-disk magnetogram at solar minimum on October 12, 1996. There are no sunspots or active regions at this time. Compared with Fig. 2, the sensitivity is increased to better show the weak magnetic features. These are either old facular remnants or short-lived ephemeral regions. Notice a weak grouping of white (north polarity) regions near the north heliographic pole (top) and of black (south polarity) regions at the south (bottom). These correspond to what is called polar faculae, a phenomenon dependent on the phase of cycle. (B) A white light solar image.

If spectroscopic observations are made in unimaged sunlight, in effect one is observing the sun as a star. Full-disk indexes include spectrum line strengths, or magnetic fields summed over the disk, or simply a number depending on the total count of spots on the disk. Full-disk solar indexes are found to track the activity cycle very well and can serve as surrogates for total solar irradiance (TSI). Particularly if one averages in time, solar indexes give nearly identical-looking signals. Figure 4 compares TSI, sunspot number, chromospheric He (1083 nm), chromospheric CaK, and the magnetic field obtained by averaging, without regard to polarity, the magnetic field over the whole disk. (If one pays attention to polarity, the + north-seeking fields largely cancel out the –south-polarity fields and the field average approaches zero.) Most of the differences seen in Fig. 4 arise from incompleteness of the various index time series owing to observing conditions (weather) or other factors. This means that long-term records, such as sunspot number, allow

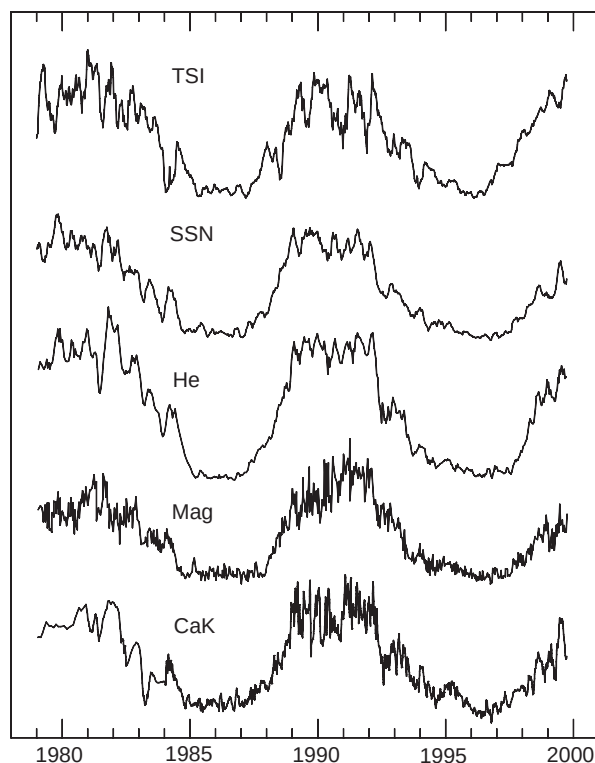


FIGURE 4 Full-disk surrogates for solar activity. TSI, total solar irradiance composite record according to Frohlich and Lean; SSN, international sunspot number; He, helium (1083-nm) line depth; Mag, magnetic field measured with the Vacuum Telescope on Kitt Peak; CaK, calcium K line central intensity measured at Kitt Peak and at Sacramento Peak. A running mean has been applied according to the completeness of the data.

one to extrapolate recent TSI measurements, which began only in 1978, back to earlier epochs.

Historic records of sunspot number, since the discovery of sunspots by Galileo, have shown intervals where there are markedly few sunspots. Named after their chroniclers, these epochs are called the Maunder (1645–1715) and the Dalton minima (1790–1820) (Fig. 5). Besides the 11-year Schwabe cycle, there is a weaker 90-year so-called Gleisberg cycle. Other cycles emerge from indirect indexes of radioisotopes, such as ^{10}Be and ^{14}C , in terrestrial reservoirs.

3. TOTAL SOLAR IRRADIANCE

Table I summarizes the energy output of the sun. By far, electromagnetic radiation (light) dominates. This is called TSI, which excludes a minute contribution at radio wavelengths and, due to variable transmission of the earth's atmosphere, must be measured with spacecraft radiometers. Presumably, the neutrino flux does not count because it is not absorbed on earth. The energy distribution by wavelength is given in Table II.

The modern radiometer is a precision device represented by the Active Cavity Radiometer Irradiance Monitor (ACRIM). ACRIM has been placed on-board a succession of satellites around the earth. It consists of a black body cavity into which sunlight is admitted through an exactly known aperture. When a cover is opened, the temperature rise over a time interval is compared with the comparable temperature rise due to electrical heating in an identical but closed comparison cavity. Once corrected for departure of the satellite from 1 AU, precise measurement of this electrical quantity of heat yields the TSI (in W/m^2). Both cavities are connected to heat sinks that periodically dump their accumulated heat.

Errors in TSI are determined by preflight calibration uncertainty and some deterioration of the black cavity due to outgassing and ultraviolet (UV) radiation damage in space. Two systems are always flown, with one being exposed only occasionally. Despite all precautions, it has proved to be difficult to cross-calibrate radiometers with sufficient accuracy. It has been proposed that radiometers be placed on the International Space Station. This would allow periodic recalibrations from the ground. The results from several radiometer experiments are summarized in Fig. 6, where the following is concluded:

- TSI is $1367 \text{ W}/\text{m}^2 \pm 0.1\%$.
- Accuracy (absolute calibration) is approximately 0.1%.
- Precision (scatter or noise) is 0.01%.

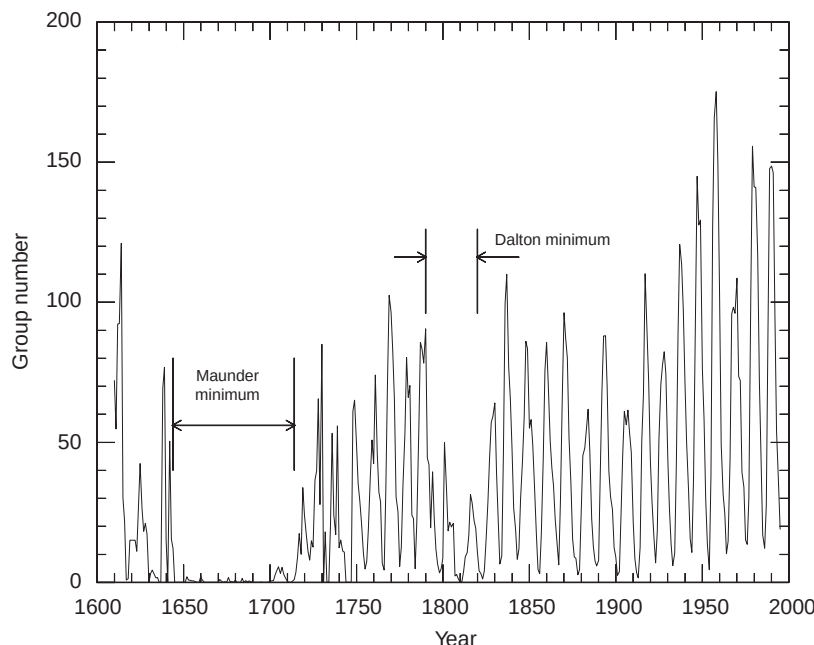


FIGURE 5 Group sunspot numbers clearly displaying the Maunder and Dalton minima.

TABLE I

Total Energy (W/m^2 at 1 AU)

Source	Energy	Activity cycle variation (percentage)
Radiation		
Irradiance	1367.10	0.08
Particle and fields		
Neutrinos	50	50 (?)
Solarwind	10^{-3}	None (?)
Magnetic field	10^{-5}	None (?)

Most of the TSI variance displayed in Fig. 6, within any given experiment, arises not from errors but rather from true solar variability imposed by dark sunspots (-0.1%) and bright faculae ($+0.1\%$). These features are seen on the solar disk in Fig. 7. Both sunspots and faculae are magnetic features; thus, the magnetic activity cycle is revealed. Careful photometry of the solar disk by astronomers at California State University, Northridge, has demonstrated that 90% of the TSI variance can be accounted for by sunspots and faculae. New results from the high-resolution Swedish Telescope indicate that perhaps tiny, not previously seen, faculae may make up part of the remaining 10%.

TABLE II

Distribution of Energy (W/m^2 at 1 AU)

Wavelength	Energy	Activity cycle variation (percentage)
$4\mu\text{m to } \infty$	0.7	—
300 nm to $4\mu\text{m}$	1350	0.08
120 to 300 nm	16	10–50
Lyman- α	3×10^{-3}	100
30 to 120 nm	2×10^{-3}	150
10 to 30 nm	10^{-5}	—

When experts model the radiometer experiments and attempt to connect the various experiments, two different time series emerge. One, by Claus Frohlich and Judith Lean, is seen in the upper part of Fig. 8. This record displays no trends beyond the activity cycle. The second and lower plot, by Richard C. Willson and Alex Mordvinov, displays increased irradiance ($+0.04\%$) at the 1995–1996 minimum. Curiously, the latter finds possible supporting evidence from the behavior of Fraunhofer line strengths, as observed with the 1-m Fourier Transform Spectrometer (FTS) at Kitt Peak in Arizona (Fig. 9). Besides the FTS data, David Gray has studied the behavior of the weak temperature-sensitive C (538-nm) line, which is formed at practically the same level as the white light continuum giving rise to the TSI signal. Gray found an increase in

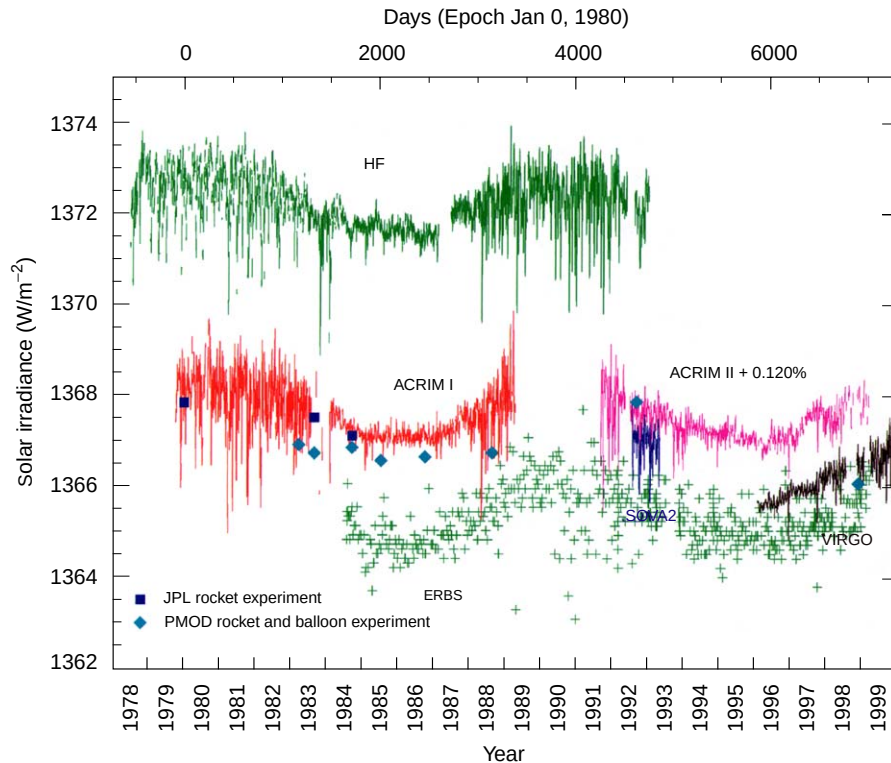


FIGURE 6 Raw total solar irradiance data, uncorrected except for adjustment to 1 AU. HF, Hickey–Frieden experiment on Nimbus 7; ACRIM, Active Cavity Radiation Irradiance Monitor on Solar Max Mission (SMM) and later on Earth Radiation Budget Satellite (ERBS); SOVA2, a solar variability experiment that has flown on several short-term flights; VIRGO, Variability of Solar Irradiance and Gravity Oscillations experiment on the SOHO spacecraft. Results from several rocket and balloon flights are included. The latter offer the advantage of postflight calibrations.

photospheric temperature of 0.014 K/year. These various, and in some ways discordant, results remain open to further study at this time.

A new set of radiometers was launched on the Solar Radiation and Climate Experiment (SORCE) early in 2003. The aim of this experiment is complete coverage of the solar output from X rays to 2.7 μm in the infrared. This data had not been released at the time of this writing. A proposed instrumental advance could be “cryogenic radiometers,” but such devices have not yet been tested in space.

One might expect the quiet sun atmosphere to be heated by faculae or perhaps cooled by sunspots. Full-disk magnetograms (i.e., pictures that map the sun’s magnetism) look very different at activity maximum from how they look at activity minimum (Figs. 2 and 3). At maximum, much of the solar surface seems to be covered by magnetic fields, but such pictures are misleading for two reasons: insufficient resolution and the fact that the fields expand above the photosphere, where the measurements are made, so as to appear more contiguous than they really are. Recent high-resolution pictures show that fields at the photo-

spheric surface are in fact sparse compared with nonmagnetic surface area. Figure 10, taken in an active region near the disk center and at much higher resolution than either the facular picture in Fig. 6 or the magnetograms, illustrates this. It is not quite a picture in white light; rather, it was taken in the “G-band” (i.e., a narrow blue filter centered at 430.4 nm), which favors the middle photosphere level. This allows one to see faculae at the disk center. The small bright wiggles are really faculae that are cospatial with the magnetic fields on magnetograms. Notice that these faculae largely avoid the granules. Time-series “movies” show that the granules push around the faculae but are otherwise unaffected by them. In other words, magnetic fields thread through the solar surface without any apparent interaction with the gas there. Temperature measurements (using temperature-sensitive spectrum lines) show that the quiet sun gas is in fact the same temperature at solar minimum as at solar maximum. This means the apparent TSI rise at solar maximum represents magnetic features only and not temperature change of the global sun. (The previously cited results of Willson

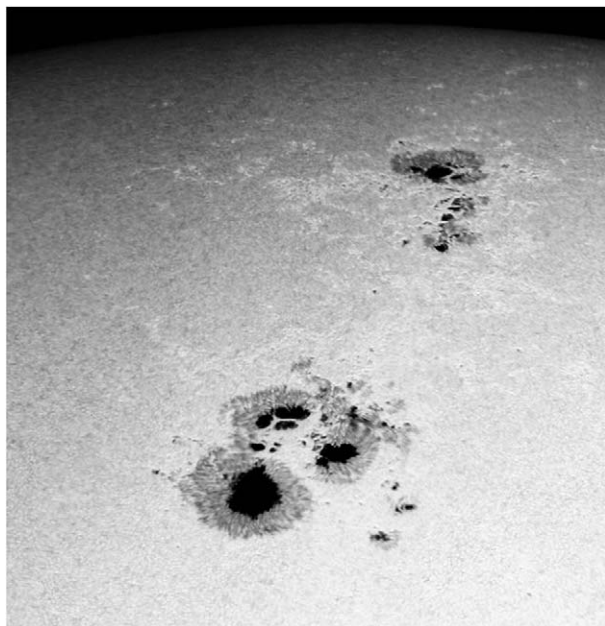


FIGURE 7 Photograph showing apparent sources of TSI variation that are evident near the solar limb. Dark sunspot umbrae, with surrounding fibril penumbrae, suppress radiation. Bright faculae, which are longer lived than sunspots, cause a net rise of TSI. Limb darkening, or the decrease in white light disk intensity as one approaches the limb, is not known to vary with time.

and Mordvinov, and of Gray, of a temperature rise minute to minute runs counter to this statement, and the matter requires further clarification.)

Theory supports these observational results. One might expect to see a bright ring around a sunspot to balance the energy trapped at the spot umbra. Henk Spruit modeled this condition and found that the energy diffuses into the convective envelope with a time constant of 1 month to 1 year. There is no localized increase in temperature. (Reported bright rings around spots turn out to be indications of abnormal granulation that is pushed around by facular fields.)

If the sun's temperature is constant, maximum to minimum, where does the 0.1% excess energy at solar maximum come from? Eugene Parker, eminent theorist of the "solar wind" and "solar dynamo," concluded that at this time we do not understand the exact physics of a variable-output sun.

Although the sun's radiative output in the UV is very small (Table II), emission lines such as Lyman- α display a large cycle variability and affect the earth's thermosphere. During solar maximum, there can be a threefold increase in the atmospheric density at an altitude of 600 km. This induces noticeable drag to low earth orbit satellites such as the Hubble Space Telescope, so they must be reboosted from time to time.

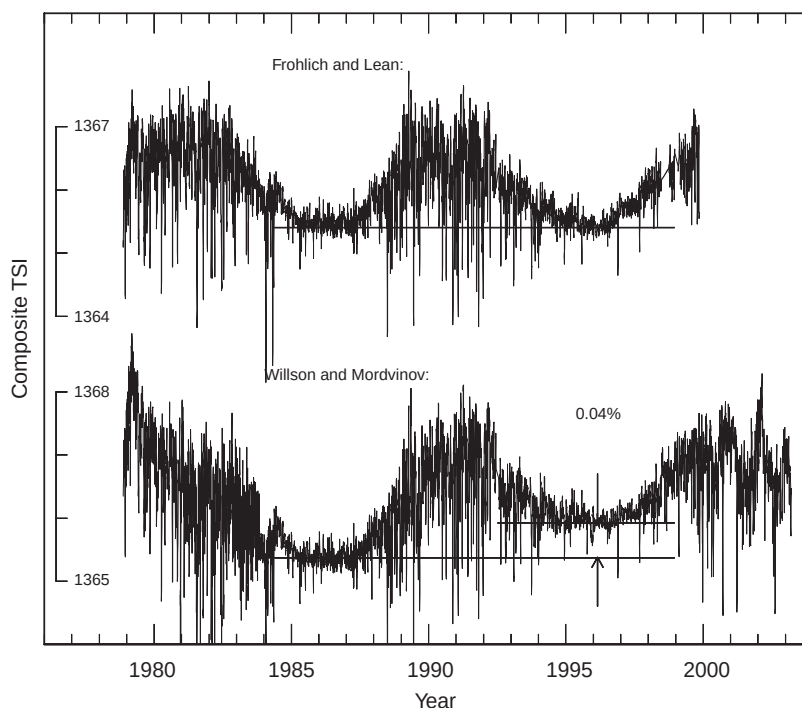


FIGURE 8 Same data as in Fig. 6 modeled in an attempt to provide a composite record capable of showing longer time trends. The Frohlich and Lean result at top shows the activity cycle with no trends. The Willson and Mordvinov construction at bottom displays a 0.04% rise between solar minima in 1985–1987 and 1995–1997.

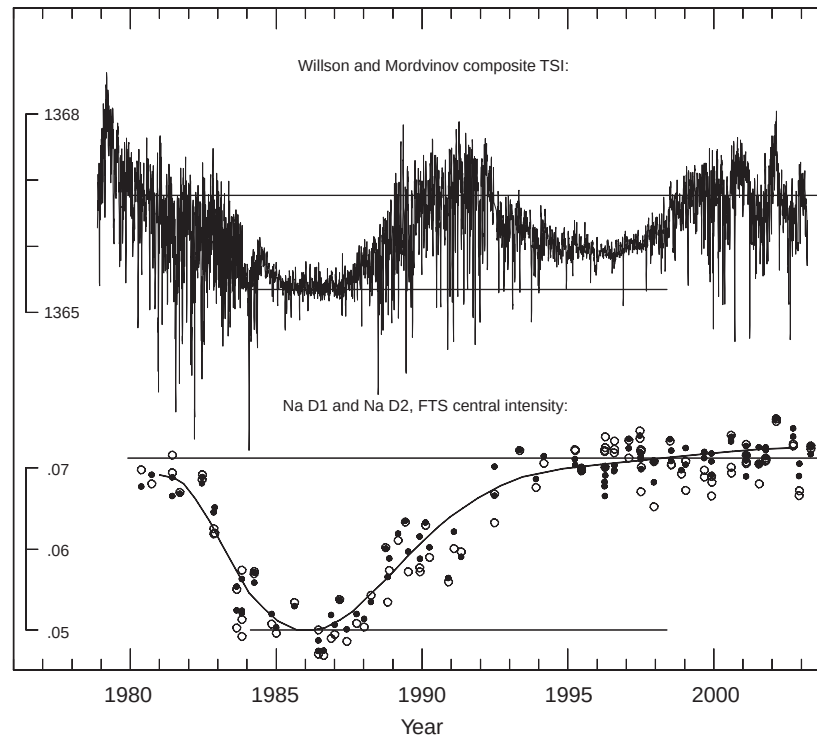


FIGURE 9 Comparison of the Willson and Mordvinov TSI solar minimum rise with the behavior of the Fraunhofer lines formed in the high photosphere. These lines failed to show the 1995–1996 minimum seen clearly in chromospheric lines. Plotted here is the central intensity of the strong sodium D lines. Other strong high photosphere lines, such as Fe and Mg, behave about the same. This indicates a temperature rise in the high photosphere at the 1996 minimum as compared with the 1986 minimum.

4. SOLAR WIND AND PARTICLE FLUX

Charged particles continually spiral outward from the sun to interact with the earth's magnetosphere. This solar wind, which is part of the extended solar atmosphere or heliosphere, traces out an Archimedes spiral path defined by the interplanetary magnetic field. Table III lists the bulk properties of the solar wind. One might expect that the chemical composition of the wind would mimic that of the photosphere, but it does not. Elements whose first ionization potential are below 10 eV are sometimes enhanced by a factor of roughly four. Some sort of selection mechanism takes place, probably in the chromosphere.

Pictures of the upper solar atmosphere, the transition region, indicate that the majority of surface fields are closed bipolar loops. However, surface evolution of magnetism may result in sizable areas where the field lines open out. These areas are called “corona holes” and are the sites of prolonged steady particle flows. When positioned favorably on the disk, coronal holes are identified with geomag-

netic storms. Most easily seen in X rays and extreme ultraviolet (EUV) pictures, such as in Fig. 11, corona holes can also be mapped from the ground using the intensity of the He (1083-nm) line.

Besides the solar wind, coronal mass ejections (CMEs) and short-lived chromospheric flares can disturb the earth environment. Figure 12 shows a large CME in progress. The frequency of CMEs varies with the solar cycle, being weekly at solar minimum and increasing to once or twice daily at solar maximum. The fact that CME activity persists at solar minimum (when the TSI is constant) is evidence that there is more to the sun's output than is revealed purely by disk radiometry.

Solar flares are a result of unstable magnetic conditions, usually associated with sunspots. Observationally, a solar flare is seen in the H- α line as a brightness increase that lasts from a few minutes up to an hour. Gamma rays, hard X rays, extreme UV radiation, and microwave radio emission are released, mostly during the “flash phase” that lasts only a few minutes. This electromagnetic radiation reaches the earth at the speed of light in 8

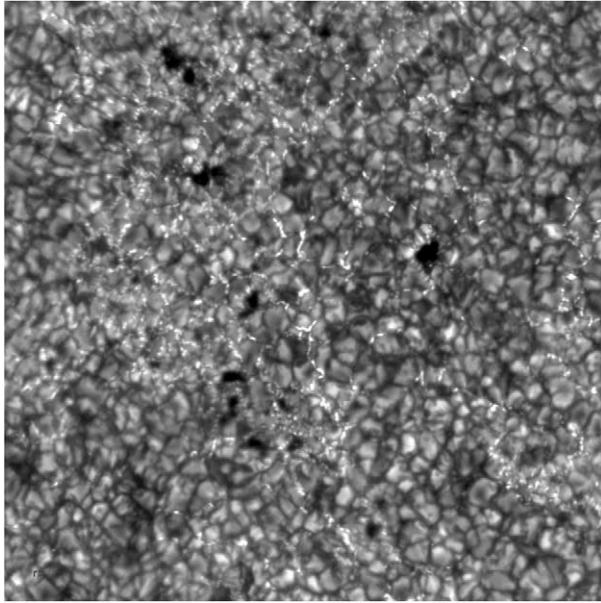


FIGURE 10 High-resolution photograph in an active region near disk center. The mottles are solar granulation similar to that in Fig. 1. Note that some granules are bright and others are not so bright. A granule is surrounded by a dark area (intergranule lane). The tiny dark features are pores, which are small sunspots without penumbrae. The bright squiggles are called filigree and represent the site of nonsunspot fields or faculae. Notice that the filigree favor intergranule lanes. Even though this is part of an active region, the relative area occupied by magnetic fields is small, and the fields thread through the granules with little or no interaction (Swedish Telescope, La Palma, Canary Islands, September 20, 1993).

TABLE III

Bulk Properties of the Solar Wind

	Low speed	High speed	Average
Particle number density N (cm^{-3})	12	4	9
Particle velocity V (km/s)	327	702	468
Particle energy (W/m^2)			1.68×10^{-3}
2.15×10^{-3}			2.05×10^{-3}
Composition (N_{α}/N_p)	0.038	0.048	0.047
Magnetic field B (y)	—	—	6.2 ± 2.9
Magnetic energy density $B^2/8\pi$ (W/m^2)	—	—	1.7×10^{-13}
Magnetic energy flux $V(B^2/8\pi)$ (W/m^2)	5.6×10^{-6}	12×10^{-6}	8×10^{-6}

min. Particles in the form of protons and alpha particles are also released in abundance, but like solar wind, they are confined to the Archimedes spiral of the interplanetary field and may take a day

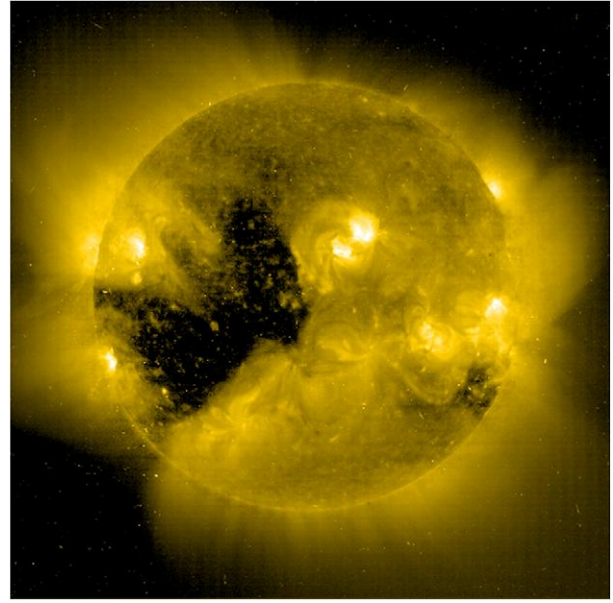


FIGURE 11 A coronal hole clearly displayed in EUV picture from the SOHO satellite on June 24, 2003. The narrow-band filter employed here is centered on Fe XV 284 Å. When this hole rotates to the western side of the sun a week later, particles from the hole will spiral out to the earth and cause geomagnetic storms.

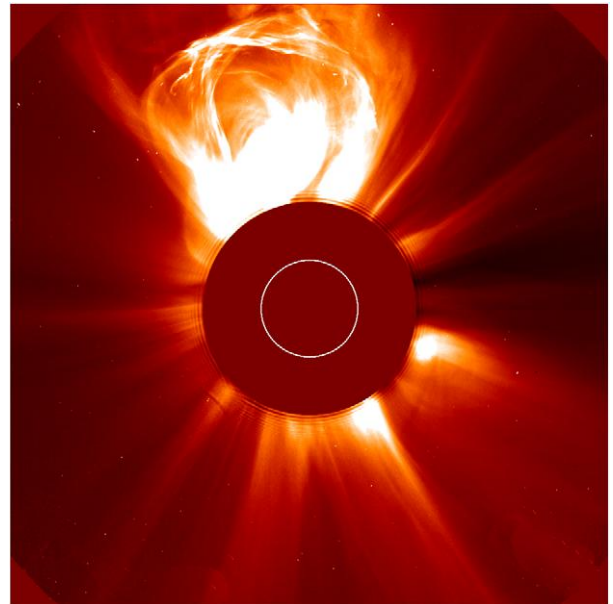


FIGURE 12 Coronal mass ejection (CME) observed on February 17, 2000, in white light by the Large Angle Spectrographic Coronagraph (LASCO) experiment on the SOHO satellite. This eruption is in the plane of the sky and not toward the earth. The solar disk is occulted with the position and size of the actual disk marked. In addition to the CME bubble, one can see the coronal streamers that are typical of solar maximum. The field of view is $6 R_{\odot}$.

or more to reach the earth. Because of their spiral path, a flare site 50° west of the meridian on the solar disk is more propitious for producing a terrestrial interaction.

5. ORBITAL MODULATION

Table IV lists geometric factors related to the energy input to the earth. Apart from the 100% modulation of the night/day diurnal cycle, the first item is solar rotation. The sun rotates with a period of 27 days in a sidereal reference frame or 25 days synodic as seen from the earth. Active regions are not distributed uniformly around the sun, and as they transit the visible disk, because of solar rotation, they modulate the TSI ($\pm 0.1\%$), $\pm 1\%$ in the near UV, and $\geq 25\%$ at Lyman- α .

The second item, and a fundamental consideration, is the variable earth-sun distance owing to the earth's elliptical orbit. This induces a 365-day modulation of $\pm 1.7\%$ in apparent TSI. Maximum energy occurs around January 4 (perihelion), whereas minimum energy occurs around July 7 (aphelion). The time of perihelion precesses with a 26,000-year period.

The third effect is more subtle and is caused by the inclination of the sun's equator to the ecliptic, or

TABLE IV
Geometric Effects

Effect	Period	Modulation amplitude
Solar rotation	27 days	λ -dependent exponential decrease with λ in UV 1% at λ 300 nm 25% at λ 120 nm
Earth-sun distance	365 days	$\pm 1.7\%$ (maximum on January 2) Precesses through the year, with 26,000 year period
Inclination of sun's equator to ecliptic	365 days	$B_0 = 0^\circ$ January 5 and June 6 $B_0 = -7.2^\circ$ March 7 $B_0 = +7.2^\circ$ September 8
Milankovich effect	40,400 years	1.2% low latitudes 2.5% high latitudes
Archimedes spiral	—	Optimum solar long: 50° west of central meridian Delay 1 to 4 days
Barycenter motion	178.7 years	?

plane, of the earth's orbit. The inclination is zero on January 5 and June 6. It is a maximum of $+7.2^\circ$ on September 8 and -7.2° on March 7. Its importance depends on the way in which active regions are distributed in latitude over the solar disk.

The fourth item is the Milankovich effect. In 1920, Milankovich, a Yugoslav astronomer, calculated that, owing to planetary perturbations; eccentricity of the earth's orbit varies with a period of approximately 100,000 years, obliquity oscillates with a period of 41,000 years, and the spin axis precesses with a period of 23,000 years. These factors, which are hard to quantify exactly, obviously influence solar insolation and may play a role in producing ice ages.

The fifth item, recently resurrected by astronomer Willie Soon, has to do with the wobble of the sun around the center of mass (barycenter) of the planetary system. In 1965, P. D. Jose pointed out that this motion ranges from 0.01 to $2.19 R_\odot$, with a period of 178.7 years. Correlations with sunspot numbers (always suspect) have been pointed out, but a physical basis for such a motion affecting sunspots is lacking.

6. LONG-TERM VARIATIONS IN SOLAR OUTPUT

This section first looks at what stellar evolution theory tells us about the past and future of the sun on cosmic time scales. Then, it examines the sunspot number record for evidence of solar variability in our own lifetimes and somewhat earlier. It extends the record back many millennia, using radioisotopes as tracers of activity cycles. Finally, it inquires about what other solar-type stars say about cycles in general and how this relates to the sun.

When the sun collapsed out of the solar nebulae, gravity was balanced by pressure, and possibly by magnetic fields, until the sun evolved into a stable "main sequence" star some 5×10^9 years ago. Initially, hydrogen burning was confined to the sun's extreme core, and its luminosity was then 70% of its current value. If the sun suddenly assumed this reduced output, the earth would freeze over. There is little or no geological evidence that this has happened in the past, and this early dim sun is considered a puzzle. Presumably the explanation has to do with the composition of the earth's primeval atmosphere, which was completely different from what it is today.

Currently, hydrogen burning in the p-p reaction is confined to within $0.3 R_\odot$. As hydrogen is further depleted after another 5×10^9 years, hydrogen burning will be confined to a shell of increasing radius and

the sun will expand greatly by a factor of 30, becoming a luminous red giant. Life will no longer be possible on earth. In fact, recent modeling suggests that the earth will become uninhabitable in only 500 million years. After the hydrogen is gone, the sun will contract, burn helium, and eventually end up a white dwarf the size of the earth. That is the big picture.

Of more interest today is solar variability at the current time and in the immediate past. In 1976, Jack Eddy compared the historic sunspot number record with auroral sightings, and with terrestrial reservoirs of ^{14}C time series, to identify the Maunder minimum. Of course, this was pre-ACRIM, but his finding of a correlation between solar and terrestrial events established the validity of the exercise.

Sunspot number archives are by no means of uniform quality over time. Since the international geophysical year (1956–1957), the competition has been keen to observe and report on sunspot numbers. But before that, the efforts were mixed, and during the century following Galileo's telescopic discovery of sunspots, we have little idea how intensively the sun was observed. French astronomers, particularly scientists such as Elizabeth Ribes at Meudon, have searched obscure sources for missing sunspot data. Recently, Doug Hoyt and Ken Schatten found that by plotting sunspot group numbers rather than simply numbers of sunspots, the early record became more complete and possibly more reliable (Fig. 5). However, there is little doubt as to the reality of the Maunder minimum.

The historic sunspot record extends back roughly 400 years, whereas radioisotope tracers can, in principle, yield information on solar activity to 100,000 BP (before present). Following a recent review article by Jürg Beer, this article now examines the past millennia of solar activity using radioisotopes found in tree rings and glacial ice as tracers.

Galactic cosmic rays flow through the solar system continually. So far as is known, this flux of cosmic rays is constant. Galactic cosmic rays are made up of protons (87%), helium nuclei (12%), and heavier elements (1%). Depending on energy, these particles interact with nitrogen and oxygen in the atmosphere to produce ^{14}C and ^{10}Be . Disturbances of the earth's geomagnetic field, and the magnetic field of the heliosphere, scatter cosmic rays and reduce their terrestrial flux. In particular, coronal mass ejections, which are associated with sunspot activity, modulate the flux of the low-energy component of cosmic rays.

Naturally occurring isotopes are ^{12}C and ^9Be , whereas the cosmic ray products are the radioisotopes ^{14}C and ^{10}Be . ^{14}C has a lifetime of 5730 years and a

cosmic ray production rate in the atmosphere of 2 atoms/cm²/s. ^{10}Be has a lifetime of 1.5×10^6 years but a production rate of only 2% that of ^{14}C .

After creation, ^{14}C undergoes a rather complex mixing process, being circulated between the atmosphere and oceans before finally being absorbed by the biosphere (e.g., in trees). For this reason, its ability to show the 11-year Schwabe cycle is attenuated by a factor of 100 and tree rings do not show the 11-year cycle. The isotope ratio $^{14}\text{C}/^{12}\text{C}$ displays up to 10% variation over the past 10,000 years, with little latitude dependence. Long-term trends are thought to be due mostly to slow changes in the geomagnetic field, with a weak but clear solar modulation superimposed to mark the Maunder minimum.

^{10}Be has a much more direct path from its source of origin in the stratosphere until, for example, deposition in snow on a Greenland glacier. Its time resolution is roughly 1 year, so that the Schwabe cycle is nicely revealed. In common with ^{14}C , it is modulated mainly by the geomagnetic field. It shows a decided preference for polar latitudes, whereas ^{14}C is latitude independent. These differences between ^{14}C and ^{10}Be can be used to cancel the dominant signal from geomagnetism and geochemical effects. Anthropogenic change from carbon-burning fuels is a new and disturbing signal during the 19th century.

A spectral analysis by Paul Damon and colleagues of the ^{14}C tree ring records revealed the 88-year Gleissberg cycle. Similar study of ^{10}Be ice core records showed the 11-year Schwabe cycle and a 205-year cycle that is named after De Vries. Both ^{14}C and ^{10}Be records show periods of low activity that are designated Maunder, Spoerer, and Wolf minima. Although dating is a potential problem, ^{10}Be ice cores can trace the Schwabe cycle back 100,000 years.

Finally, this article turns to solar-type stars as a guide to activity cycles and what to expect. The sun is classified as a G2V (i.e., a main sequence dwarf), based mainly on the relative strengths of its spectrum lines. (Astronomers classify stars as O, B, A, F, G, K, or M, based on low dispersion spectra. The series represents a run of temperature, with O being hot stars with He II absorption and M being cool stars with strong molecular TiO bands. Each class is further subdivided by numbered category (1, 2, ...9). In addition, there are luminosity classes ranging from I (supergiants) to V (main sequence dwarfs). An experienced observer can immediately classify a star simply by inspection of its spectrum. The photometric color of stars is a rough guide to spectral type and is a handy method when large numbers of stars must be examined. Of course, spectrum study requires spectra.)

During the 1960s, Olin Wilson began a program of intensive observations of the calcium H and K lines in stars judged to be similar to the sun. This was undertaken at Mount Wilson Observatory in Southern California, and many well-known astronomers, notably Sallie Baliunas and Robert Noyes, joined in the effort. The idea was an exciting one. Are there stars out there just like our sun? Of course, there are vast numbers of stars to be considered. The first criterion was blue-visual magnitude (B-V) or color (for the sun, $B-V=0.65$), then spectral class as described previously. Finally, there is the question of age gleaned from the spectrum. It is known that activity cycles decline with age. Some idea about activity, and hence age, is inferred from the strength of the calcium H and K lines.

One of the first results of this search was a curious dearth of stars like the sun—solar twins. Even to this day, and after studying hundreds of candidates, no exact twin has been found. At the time of this writing, the closest solar-like star is 18 Scopii, whose spectrum is roughly the same but whose cycle activity has a greater amplitude.

On the positive side, many cycles are found that are comparable to that of the sun. Other stars, which otherwise qualify as to age and spectrum, exhibit little or no cyclic amplitude. It is conjectured that these stars may be in a state comparable to the Maunder minimum. Looking back on the millennia records from ^{14}C , one will notice that the sun has been inactive every few hundred years or roughly one-third of the time. This condition could account for stars without cycles, and it supports the hypothesis that the sun may again become inactive. On the other hand, depending on circumstances at the birth of a star, there may be cases where the nebular condensation had very low angular momentum initially. This would create stars without appreciable activity cycles, that is, stars in a perpetual state of Maunder minimum.

The question of age is a complication circumvented by a scheme devised by Mark Giampapa and Richard Radick. With help from fellow astronomers, they have been observing solar-type stars in the open-cluster M67. The advantage is that all of the stars would be expected to have the same age, thereby eliminating that variable. What has been found so far

is that solar-type cycles do dominate but that there seems to be a separate class of extra-strong cycles. This project has been ongoing.

The ^{10}Be archives strongly suggest, from signals of solar-type stars and from solar activity such as CMEs (which continue to bombard the earth even at solar minimum), that there is a lot going on with the sun of which we are currently unaware. Plenty of work exists for future generations of solar physicists.

Acknowledgments

The author acknowledges help in the preparation of this article from Bernard Durney, Peter Foukal, Harry Jones, Steve Keil, Elena Malaushenko, Judit Pap, and Dick White.

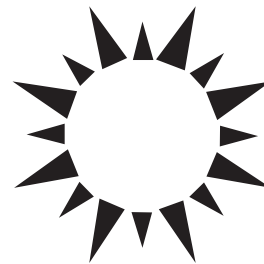
SEE ALSO THE FOLLOWING ARTICLES

Earth's Energy Balance • Global Material Cycles and Energy • Lunar-Solar Power System • Materials for Solar Energy • Origin of Life and Energy • Photosynthesis and Autotrophic Energy Flows • Solar Energy, History of • Solar Fuels and Materials

Further Reading

- Beer, J. (2000). Long-term indirect indices of solar variability. *Space Sci. Rev.* **94**, 53.
- Cox, A. N., Livingston, W. C., and Mathews, M. S. (eds.). (1991). "Solar Interior and Atmosphere." University of Arizona Press, Tucson.
- Eddy, J. A. (1976). The maunder minimum. *Science* **192**, 1189.
- Foukal, P. (1990). "Solar Astrophysics." John Wiley, New York.
- Frohlich, C., and Lean, J. (1998). The sun's total irradiance: Cycles, trends, and related climate change uncertainties since 1976. *Geophys. Res. Lett.* **25**, 4377.
- Jose, P. D. (1965). Sun's motion and sunspots. *Astron. J.* **70**, 193.
- Parker, E. N. (2000, June). The physics of the sun and the gateway to the stars. *Physics Today*, p. 26.
- Spruit, H. (1982). Effect of spots on a star's radius and luminosity. *Astron. Astrophys.* **108**, 348.
- Stix, M. (2002). "The Sun." Springer, New York.
- White, O. R., Rottman, G. J., Woods, T. N., Knapp, B. C., Keil, S. L., Livingston, W. C., Tapping, K. F., Donnelly, R. F., and Puga, L. C. (1994). Change in the radiative output of the sun in 1992 and its effect in the thermosphere. *J. Geophys. Res.* **99**, 369.

Sustainable Development: Basic Concepts and Application to Energy



MOHAN MUNASINGHE

Munasinghe Institute for Development
Colombo, Sri Lanka

-
1. Basic Framework
 2. Elements of Sustainable Development
 3. Sustainable Energy Development
 4. Applying the Sustainable Energy Development Framework

Glossary

Action Impact Matrix (AIM) A sustainomics tool that helps to identify and analyze economic–environmental–social interactions and to formulate effective sustainable development policies; individual cells in the matrix help to explicitly identify key linkages, focus attention on methods of analyzing important impacts, and suggest action priorities and remedies.

durability An integrative approach focusing primarily on sustaining the quality of life, typically by meeting prudent environmental, social, and economic sustainability requirements; durable development paths permit growth but are not necessarily economically optimal.

economic sustainability Maximizing the flow of income that could be generated while at least maintaining the stock of assets (or capital) that yields these beneficial outputs; economic efficiency continues to ensure both efficient resource allocation in production and efficient consumption choices that maximize utility.

environmental sustainability Preserving the overall viability and normal functioning of natural systems, including improving system health and dynamic ability to adapt to change across a range of spatial and temporal scales; for ecological systems, sustainability is defined by a comprehensive, multiscale, dynamic hierarchical measure of resilience, vigor, and organization.

equity An ethical and people-oriented concept focusing on the basic fairness of both the processes and outcomes of decision making, including the distribution of economic, social, and environmental assets and entitlements.

optimality An approach used in economic analysis to generally maximize (or minimize) a single objective such as welfare (or costs), subject to the requirement that the stock of productive assets (or welfare itself) is non-decreasing in the long term.

social sustainability Preserving the overall viability and normal functioning of social systems.

sustainable development A process for improving the range of opportunities that will enable individual humans and communities to achieve their aspirations and full potential over a sustained period of time while maintaining the resilience of economic, social, and environmental systems.

sustainable energy development The harnessing of energy resources for human use in a manner that supports lasting development.

sustainomics A transdisciplinary, integrative, comprehensive, balanced, heuristic, and practical meta-framework for making development more sustainable.

Following the 1992 Earth Summit in Rio de Janeiro, Brazil, including the adoption of the United Nations' Agenda 21, and the 2002 World Summit on Sustainable Development in Johannesburg, South Africa, sustainable development has become well accepted worldwide. World decision makers are looking at this approach to solve many critical problems, including traditional development issues (e.g., energy use, economic stagnation, poverty, hunger, illness) as well as newer challenges (e.g., environmental degradation, globalization).

1. BASIC FRAMEWORK

Sustainable energy development involves the harnessing of energy resources for human use in a manner

that supports lasting development. This article begins with a review of sustainable development itself before describing its application to energy. The World Commission on Environment and Development originally defined it as “development which meets the needs of the present without compromising the ability of future generations to meet their own needs.” Among many subsequent definitions, the sustainable development triangle in Fig. 1A shows one widely accepted concept proposed by Munasinghe at the 1992 Earth Summit in Rio de Janeiro, Brazil. It encompasses three major perspectives: economic, social, and environmental. Each viewpoint corresponds to a domain (and system) that has its own distinct driving forces and objectives. The economy is geared toward improving human welfare, primarily through increases in consumption of goods and services. The environmental domain focuses on protecting the integrity and resilience of ecological systems. The social domain emphasizes enriching human relationships, achieving individual and group aspirations, and strengthening values and institutions.

Meanwhile, energy has emerged as a key resource that interacts critically with the economic, social, and environmental dimensions of sustainable development. First, it has long been perceived as a major driving force underlying economic progress, and in turn, economic growth itself further stimulates energy demand. Second, energy production and use are strongly interlinked with the environment. Third, energy is a basic human need that significantly affects

social well-being. Recently, growing energy demand has also become associated with global climate change, posing an unprecedented challenge to humanity. The wide-ranging potential impacts of energy production and consumption on sustainable development, and vice versa, are reviewed in this article.

Given the lack of a specific approach or practical framework that attempts to define, analyze, and implement sustainable development, Munasinghe proposed the term “sustainomics” to describe “a transdisciplinary, integrative, comprehensive, balanced, heuristic, and practical meta-framework for making development more sustainable.” Figure 1B indicates how the sustainomics framework (or the emerging “science of sustainable development”), as well as the associated transdisciplinary knowledge base, would support a comprehensive and balanced assessment of trade-offs and synergies that might exist among the economic, social, and environmental dimensions of sustainable development. Balance is also needed in the relative emphasis placed on traditional development (which is more appealing to the South) versus sustainability (which is emphasized by the North). The optimality and durability approaches (described later) play a key role in integrating economic, social, and environmental issues. Many disciplines contribute to the sustainomics framework given that sustainable development itself involves every aspect of human activity, including complex interactions among socioeconomic, ecological, and physical systems.

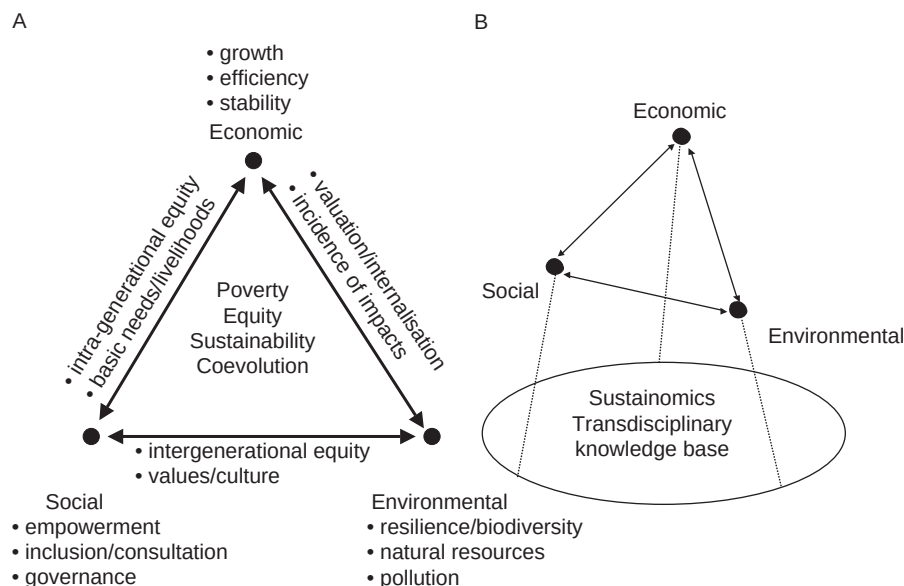


FIGURE 1 (A) Elements of sustainable development. (B) Sustainable development triangle supported by the sustainomics framework. Adapted from Munasinghe (1992).

Historically, the development of the industrialized world has focused on material production. Not surprisingly, most industrialized and developing nations pursued the economic goal of increasing output and growth during the 20th century. Thus, traditional development was strongly associated with economic growth as well as with important social dimensions (poverty and equity are discussed later). By the early 1960s, the lack of “trickle down” benefits to the growing numbers of poor in developing countries resulted in greater efforts to improve income distribution directly. The development paradigm shifted toward equitable growth, where social (distributional) objectives, especially poverty alleviation, were recognized to be as important as economic efficiency. Protection of the environment has now become the third major pillar of sustainable development. By the early 1980s, a large body of evidence had accumulated that environmental degradation was a major barrier to development, and new proactive safeguards were introduced (e.g., the environmental assessments).

In the sustainomics framework, sustainable development is described as a process for improving the range of opportunities that will enable individual humans and communities to achieve their aspirations and full potential over a sustained period of time while maintaining the resilience of economic, social, and environmental systems. Thus, sustainable development requires both increases in adaptive capacity and opportunities for improving economic, social, and ecological systems. Improving adaptive capacity increases resilience and sustainability. Expanding the set of opportunities for improvement gives rise to development. Heuristic behavior of individual organisms and systems facilitates learning, the testing of new processes, adaptation, and improvement.

The precise definition and implementation of sustainable development remains an ideal, elusive, and perhaps unreachable goal. Sustainomics proposes a less ambitious, but more focused and feasible, strategy that merely seeks to “make development more sustainable.” Such an incremental (or gradient-based) method is more practical because many unsustainable activities are easier to recognize and eliminate. This approach seeks continuing improvements in the current quality of life at a lower intensity of resource use, thereby leaving behind for future generations an undiminished stock of productive assets (e.g., manufactured, natural, and social capital) that will enhance opportunities for improving their quality of life.

2. ELEMENTS OF SUSTAINABLE DEVELOPMENT

2.1 Economic Aspects

Economic progress is often evaluated in terms of welfare (or utility), measured as willingness to pay for goods and services consumed. Thus, many economic policies typically seek to enhance income and induce more efficient production and consumption of (mainly marketed) goods and services. The stability of prices and employment are among other important objectives.

Economic efficiency is measured against the ideal of Pareto optimality, which encourages actions that will improve the welfare of at least one individual without worsening the situation of anyone else. The idealized, perfectly competitive economy is an important (Pareto optimal) benchmark, where (efficient) market prices play a key role in both allocating productive resources to maximize output and ensuring optimal consumption choices that maximize consumer utility. If significant economic distortions are present, appropriate shadow prices need to be used. The well-known cost–benefit criterion accepts all projects whose net benefits are positive (i.e., aggregate benefits exceed costs). It is based on the weaker “quasi”-Pareto condition, which assumes that such net benefits could be redistributed from potential gainers to losers, leaving no one worse off than before. In more general terms, interpersonal comparisons of welfare are fraught with difficulty, both within and across nations and over time (e.g., the value of human life).

Based on the pioneering work of Lindahl and Hicks, the modern concept underlying economic sustainability seeks to maximize the flow of income that could be generated while at least maintaining the stock of assets (or capital) that yields these beneficial outputs. Economic efficiency continues to ensure both efficient resource allocation in production and efficient consumption choices that maximize utility. Problems arise in identifying the kinds of capital to be maintained (e.g., manufactured, natural, human, and social capital) and their substitutability. Often, it is difficult to value these assets and the services they provide, particularly for ecological and social resources. Even key economic assets may be overlooked, especially in informal or subsistence economies where non-market-based transactions are important. Meanwhile, the equation of welfare with monetary income and consumption has been challenged for many years. More recently, Maslow and

others have identified hierarchies of needs that provide psychic satisfaction beyond mere goods and services.

The issues of uncertainty, irreversibility, and catastrophic collapse pose additional difficulties in determining dynamically efficient development paths. Many commonly used microeconomic approaches rely on marginal analysis (e.g., comparing incremental costs and benefits of economic activities). Such methods assume smoothly changing variables and, therefore, are rather inappropriate for analyzing large changes, discontinuous phenomena, and sudden transitions among multiple equilibria. Recent work has begun to explore the behavior of large, nonlinear, dynamic, and chaotic systems and concepts such as system vulnerability and resilience.

2.2 Environmental Aspects

Development in the environmental sense is a recent concern relating to the need to manage scarce natural resources in a prudent manner because human welfare ultimately depends on ecological services. Ignoring safe ecological limits could undermine long-run prospects for development. The growing literature on the theme of environment and sustainable development includes books by Faucheux and colleagues, describing models of sustainable development, and Munasinghe and colleagues, explicitly addressing the links between growth and environment

The environmental interpretation of sustainability focuses on overall viability and normal functioning of natural systems. For ecological systems, sustainability is defined by a comprehensive, multiscale, dynamic hierarchical measure of resilience, vigor, and organization. In 1973, Holling defined resilience as the ability of an ecosystem to persist despite external shocks, that is, the amount of disruption that will cause the ecosystem to switch from one system state to another. An ecosystem state is defined by its internal structure and set of mutually reinforcing processes. Vigor is associated with the primary productivity or growth of an ecosystem. Organization depends on both complexity and structure of the system. For example, a multicellular organism such as a human is more highly organized than a single-celled amoeba. Higher states of organization imply lower levels of entropy. Thus, the second law of thermodynamics requires that sustainability of complex organisms and systems depends on the use of low-entropy energy derived from their environment that is returned as (less useful) high-entropy energy. The ultimate source of this energy is solar radiation.

In this context, natural resource degradation, pollution, and loss of biodiversity are detrimental because they increase vulnerability, undermine system health, and reduce resilience. The notions of a safe threshold and carrying capacity are important, often to avoid catastrophic ecosystem collapse. It is useful to also think of sustainability in terms of the normal functioning and longevity of a nested hierarchy of ecological and socioeconomic systems ordered according to scale; for example, a human community would consist of many individuals who are themselves composed of a large number of discrete cells. Gunderson and Holling used the term “panarchy” to denote such a hierarchy of systems and their adaptive cycles across scales. A system at a given level is able to operate in its stable (sustainable) mode because it is protected by slower and more conservative changes in the supersystem above it while being simultaneously invigorated and energized by faster changes taking place in subsystems below it.

Sustainable development is not necessarily synonymous with maintaining the ecological status quo. A coupled ecological–socioeconomic system could evolve while maintaining levels of biodiversity that guarantee resilience of ecosystems on which future human consumption and production depend.

2.3 Social Aspects

Social development usually refers to improvements in both individual well-being and the overall welfare of society resulting from increases in social capital, typically the accumulation of capacity enabling individuals and communities to work together. The institutional component of social capital involves formal laws as well as traditional or informal understandings that govern behavior, whereas the organizational component is embodied in individuals and communities operating within these institutional arrangements. The quantity and quality of social interactions underlying human existence (including levels of mutual trust and shared social norms) help to determine the stock of social capital. Thus, social capital grows with greater use and erodes through disuse, unlike economic and environmental capital, which are depreciated or depleted by use. Furthermore, some forms of social capital may be harmful (e.g., cooperation within criminal gangs).

There is also an important element of equity and poverty alleviation. Thus, the social dimension of development includes protective strategies that reduce vulnerability, improve equity, and ensure that basic needs are met. Future social development will

require sociopolitical institutions that can adapt to meet the challenges of globalization. The latter often destroys traditional coping mechanisms that have evolved in the past (especially to protect disadvantaged groups).

Social sustainability parallels environmental sustainability because human settlements, such as cities and villages, are also habitats. Reducing vulnerability and maintaining the health (e.g., resilience, vigor, organization) of sociocultural systems and their ability to withstand shocks are also important. Enhancing human capital (through education) and strengthening social values, institutions, and governance are key aspects. Many harmful changes occur slowly, and their long-term effects are often overlooked in socioeconomic analysis. Preserving cultural capital and diversity worldwide, strengthening social cohesion, and reducing destructive conflicts are integral elements of this approach. An important aspect involves empowerment and broader participation through subsidiarity, that is, decentralization of decision making to the lowest (or most local) level at which it is still effective. In summary, for both ecological and socioeconomic systems, the emphasis is on improving system health and its dynamic ability to adapt to change across a range of spatial and temporal scales rather than on conserving some “ideal” static state.

2.4 Poverty and Equity

Poverty and equity are two important issues that have social, economic, and environmental dimensions (Fig. 1A). More than 2.8 billion people (nearly half of the global population) live on less than U.S. \$2 per day, and 1.2 billion barely survive on less than U.S. \$1 per day. The top 20 percentile of the world’s population consumes approximately 83% of total output, whereas the bottom 20 percentile consumes only 1.4%. Income disparities are worsening; the per capita ratio between the richest and the poorest 20 percentile groups was 30 to 1 in 1960 and increased to more than 80 to 1 by 1995.

Equity is an ethical and usually people-oriented concept with primarily social, and some economic and environmental, dimensions. It focuses on the basic fairness of both the processes and outcomes of decision making. The equity of any action may be assessed in terms of a number of generic approaches, including parity, proportionality, priority, utilitarianism, and Rawlsian distributive justice. Societies normally seek to achieve equity by balancing and combining several of these criteria.

Poverty alleviation, improved income distribution, and intragenerational (or spatial) equity are key aspects of economic policies seeking to increase overall human welfare. Broadly speaking, economic efficiency provides guidance on producing and consuming goods and services more efficiently but is unable to provide a means of choosing (from a social perspective) among alternative patterns of consumption that are efficient. Equity principles provide better tools for making judgments about such choices.

Social equity is also linked to sustainability because grossly unfair distributions of income and social benefits are unlikely to be lasting in the long run. Equity will be strengthened by enhancing pluralism and grassroots participation in decision making and by empowering disadvantaged groups. In the long term, intergenerational equity is vital, where the economic discount rate plays a key role with respect to both equity and efficiency aspects. The sustainomics framework outlines methods of reconciling potential conflicts between equity and economic efficiency.

Equity in the environmental sense has received recent attention due to disproportionately greater environmental damages suffered by poor groups. Thus, poverty alleviation efforts are being broadened (beyond raising monetary incomes) to address the degraded environmental and social conditions facing the poor.

In summary, both equity and poverty have not only economic but also social and environmental dimensions; therefore, they need to be assessed using a comprehensive set of indicators (rather than income distribution alone). From an economic policy perspective, emphasis needs to be placed on expanding employment and gainful opportunities for poor people through growth, improving access to markets, and increasing both assets and education. Social policies would focus on empowerment and inclusion by making institutions more responsive to the poor and removing barriers that exclude disadvantaged groups. Environmentally related measures to help poor people might seek to reduce their vulnerability to disasters, crop failures, loss of employment, sickness, economic shocks, and so on. Thus, an important objective of poverty alleviation is to provide poor people with assets (e.g., social, natural, economic) that will reduce their vulnerability. Such assets increase the capacity for both short-run coping and longer run adaptation to external shocks. The forgoing ideas blend with the sustainable livelihoods approach, which focuses on access to portfolios of

assets, capacity to withstand shocks, gainful employment, and social processes.

An even broader nonanthropocentric approach to equity involves the concept of fairness in the treatment of nonhuman forms of life or even inanimate nature. One view asserts that humans have the responsibility of prudent “stewardship” over nature, and this goes beyond mere rights of usage.

2.5 Consistent Integration of Economic, Social, and Environmental Considerations

This subsection compares the concepts of ecological, social, and economic sustainability. Maintaining the set of opportunities, as opposed to preserving value of the asset base, is important. The preservation of biodiversity maintains options and allows the system to retain resilience by protecting it from external shocks, in the same manner that preservation of the capital stock protects economic assets for future consumption. Differences emerge because under the Hicks–Lindahl income measure, a society that consumes its fixed capital without replacement is not sustainable, whereas using an ecological approach, loss of resilience implies a reduction in the self-organization of the system but not necessarily a loss in productivity. In the case of social systems, resilience depends on the capacity of human societies to adapt and continue functioning in the face of stress and shocks. Thus, linkages between socio-cultural and ecological sustainability emerge through the organizational similarities between human societies and ecological systems and the parallels between biodiversity and cultural diversity. In the longer term, the concept of coevolution of social, economic, and ecological systems within a larger, more complex adaptive system provides useful insights regarding harmonious integration of the various elements of sustainable development (Fig. 1A).

It is important to integrate and reconcile the economic, social, and environmental aspects within a holistic and balanced sustainable development framework. Economic analysis has a special role in contemporary national policymaking because some of the most important decisions fall within the economic domain. Although mainstream economics that is used for practical policymaking has often ignored many crucial aspects of the environmental and social dimensions of sustainable development, there is a small but growing body of literature that seeks to address such shortcomings.

Two broad approaches, based on the concepts of optimality and durability, are relevant for integrating

the economic, social, and environmental dimensions of sustainable development. Although there are overlaps between the two approaches, the main thrust is somewhat different in each case. Uncertainty often plays a key role in determining which approach would be preferred. For example, relatively steady and well-ordered conditions may encourage optimizing behavior by planners attempting to control and even fine-tune outcomes, whereas subsistence farmers facing chaotic and unpredictable circumstances might opt for more durable responses that simply enhance survival prospects.

2.6 Optimality

The optimality-based approach has been widely used in economic analysis to generally maximize welfare (or utility), subject to the requirement that the stock of productive assets (or welfare itself) is nondecreasing in the long term. This assumption is common to most sustainable economic growth models. The essence of the approach is illustrated by the simple example of maximization of the flow of aggregate welfare (W), cumulatively discounted over infinite time (t), as represented by the following expression:

$$\max \int_0^{\infty} W(C, Z) \cdot e^{-rt} dt.$$

Here, W is a function of C (consumption) and Z (set of other relevant variables), whereas r is the discount rate. Further side constraints may be imposed to satisfy sustainability needs (e.g., nondecreasing stocks of productive assets). One typical application of optimization involves electric generation expansion planning, where analysts seek to minimize the present discounted costs of generation to meet the desired demand forecast over several decades at a given level of reliability.

Some ecological models also optimize variables related to system vigor such as energy use, nutrient flow, and biomass production. In economic models, utility is often measured in terms of net benefits of economic activities, that is, the benefits of development activities minus corresponding costs. More sophisticated economic optimization approaches include environmental and social variables (by attempting to value environmental externalities, system resilience, etc.). However, given the difficulties of quantifying and valuing many such “noneconomic” assets, the costs and benefits associated with market-based activities tend to dominate in most economic optimization models.

Basically, the optimal growth path maximizes economic output, whereas the sustainability requirement is met by ensuring nondecreasing stocks of assets. Some analysts support a “strong sustainability” constraint, which requires separate preservation of each category of critical asset (e.g., manufactured, natural, sociocultural, and human capital), assuming that they are complements rather than substitutes. Other researchers have argued in favor of “weak sustainability,” which seeks to maintain the aggregate monetary value of the total stock of assets, assuming that various asset types are substitutes and may be valued.

Side constraints are often necessary because the optimization approach (including economic efficiency and valuation) might not be easily applied to ecological objectives, such as protecting biodiversity and improving resilience, or to social goals, such as promoting equity and empowerment. Such environmental and social variables cannot be easily incorporated within a single valued objective function based on cost–benefit analysis. Thus, techniques such as multicriteria analysis (MCA) may be required to facilitate trade-offs among noncommensurable variables. Moreover, the lagged price system might not anticipate irreversible environmental and social harm and nonlinear system responses that could lead to catastrophic collapse. Therefore, noneconomic indicators of environmental and social status would be helpful. The constraints on critical environmental and social indicators are proxies representing safe thresholds that help to maintain the viability of those systems. Risk and uncertainty will also necessitate the use of decision analysis tools.

2.7 Durability

The second broad integrative approach focuses primarily on sustaining the quality of life, for example, by satisfying environmental, social, and economic sustainability requirements. Such a framework favors “durable” development paths that permit growth but are not necessarily economically optimal. There is more willingness to trade off some economic optimality for the sake of greater safety (i.e., risk aversion) to stay within critical environmental, social, and technical limits. The durability approach would facilitate the design of energy systems that are more resilient and less vulnerable to disruptive shocks.

Economic system durability might require consumption levels to be maintained, that is, per capita consumption that never falls below some minimum

level or is nondeclining. The environmental and social sustainability requirements may be expressed in terms of indicators of “state” that monitor the durability or normal functioning of complex ecological, social, and technoeconomic systems. For example, consider an electric power system where durability may be represented by various indexes of system reliability (e.g., loss of load probability [LOLP], frequency and duration of outages [F&D]) and system stability (especially transmission and distribution systems) in relation to external shocks (e.g., storms) that might cause the breakdown of critical system components (e.g., generators, lines).

There is the likelihood of further interaction here due to linkages among the sustainability of social, ecological, and technoeconomic systems; for example, social disruption and conflict could exacerbate damage to both ecological and technoeconomic systems and vice versa. In fact, long-standing social norms in many traditional societies have helped to protect the environment.

Durability encourages a holistic systemic viewpoint, and this is important in sustainomics analysis. The self-organizing and internal structure of complex systems often makes “the whole more durable (and valuable) than the sum of the parts.” A narrow definition of efficiency based on marginal analysis of individual components may be misleading. For example, it is more difficult to value the integrated functional diversity in a forest ecosystem than to value the individual species of trees and animals. Therefore, the former is more likely to fall victim to market failure (as an externality). Furthermore, even where correct environmental shadow prices prevail, some analysts point out that cost minimization could lead to homogenization and consequent reductions in system diversity. Systems analysis also helps to identify the benefits of cooperative structures and behavior that a more partial analysis might neglect.

The possibility of many durable paths favors simulation-based methods, including consideration of alternative futures (rather than one optimal result). This approach parallels research on integrating human actors into ecological models, including multiple-agent modeling to account for heterogeneous behavior, bounded rationality leading to different perceptions and biases, and social interactions involving imitation, reciprocity, and comparison.

In the durability approach, maintaining asset stocks enhances system sustainability because various forms of capital are a bulwark that decreases vulnerability to external shocks and reduces irreversible harm rather than merely producing more

economic outputs. System vulnerability, resilience, vigor, organization, and ability to adapt will depend dynamically on the capital endowment as well as on the magnitude and rate of change of a shock; for example, in the case of an electric power system, vulnerability and risk of failure depend on the traditional reserve margin (supply minus demand) as well as on the probability of breakdown of key system components or stochastic surges in demand.

2.8 Indicators

Because asset stocks are important to both the optimal and durable approaches, the practical implementation of sustainability principles will require the identification of relevant economic, social, and environmental indicators at various levels of aggregation ranging from the global/macro to the local/micro. Indicators must be comprehensive in scope, must be multidimensional in nature (where appropriate), and must account for spatial differences.

Measuring economic, environmental, human, and social capital also raises various problems. Manufactured capital may be estimated using conventional neoclassical economic analysis. Market prices are useful when economic distortions are relatively low, whereas shadow prices could be applied in cases where market prices are unreliable. Natural capital needs to be quantified first in terms of key physical attributes. Typically, damage to natural capital may be assessed by the level of air pollution (e.g., concentrations of suspended particulates, sulfur dioxide, greenhouse gases [GHGs]), water pollution (e.g., biochemical oxygen demand [BOD], chemical oxygen demand [COD]), and land degradation (e.g., soil erosion, deforestation). Then this physical damage could be valued using environmental and resource economics techniques. Human resource stocks are often measured by educational levels, productivity, and earning potential. Social capital is more difficult to assess. Putnam described it as “horizontal associations” among people or social networks and associated behavioral norms and values that affect the productivity of communities. Social capital may be viewed more broadly in terms of social structures that facilitate the activities of agents in society, including both horizontal and vertical associations (e.g., firms). An even wider definition is implied by the institutional approach that includes not only the mainly informal relationships implied by the earlier two views but also more formal frameworks provided by governments, political systems, and legal provisions. Recent work has

sought to distinguish between social capital and political capital, that is, the networks of power and influence that link individuals and communities to the higher levels of decision making.

2.9 Complementarity and Convergence of Optimal and Durable Approaches

The two approaches are often complementary in national economic management. For example, economy-wide policies involving both fiscal and monetary measures (e.g., taxes, subsidies, interest, foreign exchange rates) might be optimized using quantitative macroeconomic models. Nevertheless, decision makers inevitably modify these economically “optimal” policies before implementing them to take into account other durable sociopolitical considerations (e.g., poverty alleviation, regional factors) that facilitate governance and stability. The determination of an appropriate target trajectory for future global GHG emissions provides another useful illustration of the interplay between durability and optimality.

In practice, the two approaches point to convergent solutions. First, wastes ought to be generated at rates within the assimilative capacity of the environment. Second, scarce renewable resources should be used at rates compatible with the natural rate of regeneration. Third, nonrenewable resource use rates should depend on the substitutability between these resources and technological progress. Both wastes and natural resource inputs might be minimized by moving from linear throughput to closed loop (or recycling) mode. Finally, inter- and intragenerational equity (especially poverty alleviation), pluralistic and consultative decision making, and enhanced social values and institutions are important additional aspects.

GHG mitigation provides an interesting example of how such an integrative framework could help to incorporate climate change policies within a national sustainable development strategy. The total GHG emissions rate (G) may be decomposed as follows:

$$G = (Q/P) \times (Y/Q) \times (G/Y) \times P,$$

where (Q/P) is quality of life per capita, (Y/Q) is material consumption required per unit of quality of life, (G/Y) is GHG emissions per unit of consumption, and P is population. A high quality of life is consistent with low total GHG emissions, provided that each of the three terms on the right-hand side could be minimized. Reducing (Y/Q) implies “social decoupling” (or “dematerialization”), whereby satisfaction becomes less dependent on material

consumption through changes in tastes, behaviors, and social values. Similarly, (G/Y) may be reduced by “technological decoupling” (or “decarbonization”) that reduces the intensity of GHG emissions in consumption and production. Finally, population growth could be reduced, especially where emissions per capita are already high. The linkages between social and technological decoupling need to be explored. For example, changes in public perceptions and tastes could affect the directions of energy technology and influence the effectiveness of mitigation and adaptation policies.

Climate change researchers are currently exploring the application of large and complex integrated assessment models (IAMs). IAMs include coupled submodels representing various ecological, geophysical, and socioeconomic systems where there is scope to apply the optimality and durability approaches consistently.

2.10 Tools for Sustainable Development Analysis and Assessment

Some important tools for sustainable development analysis and assessment include the Action Impact Matrix (AIM) for prioritizing the economic, environmental, and social interactions of various macro-economic and sectoral development policies; advanced cost-benefit analysis, including economic

valuation of environmental and social impacts; and MCA, especially in cases where some impacts cannot be easily quantified in monetary terms.

Sustainable development assessment (SDA) is another important tool to ensure balanced analysis of both development and sustainability concerns in policies and projects. The economic component of SDA is based on conventional economic and financial analysis (including cost-benefit analysis). The other two key components are environmental assessment and social assessment. SDA also includes poverty assessment. Thus, SDA seeks to integrate and harmonize economic, environmental, and social analyses.

3. SUSTAINABLE ENERGY DEVELOPMENT

3.1 Linkages between Energy Use and Sustainable Development

3.1.1 Energy-Economy Linkages

The economy-energy interaction has been well analyzed in terms of energy as a driving force for economic progress, especially since the industrial revolution and the extensive use of fossil fuels. Modern economies could not function without commercial energy. Figure 2 shows that future consumption of nearly all forms of energy will

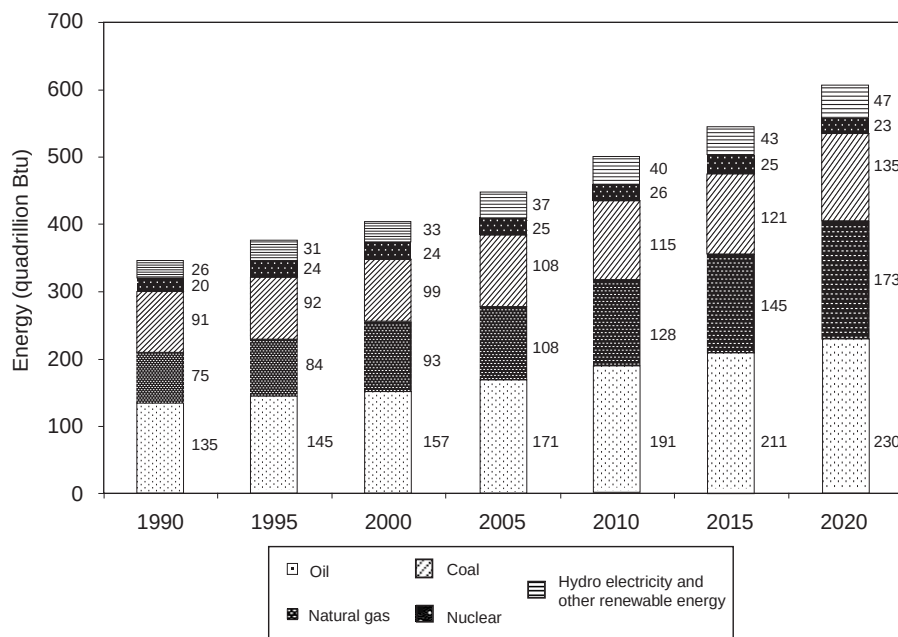


FIGURE 2 World consumption of energy by source, 1990–2020.

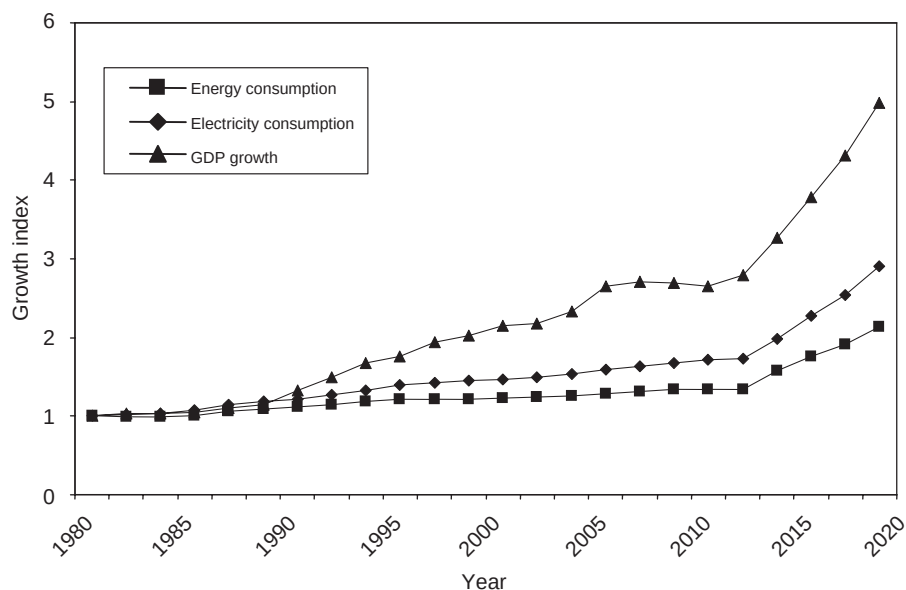


FIGURE 3 Growth of world energy and electricity consumption and gross domestic product (GDP), 1980–2020. 1980 actuals: energy, 285 quadrillion Btu; electricity, 7420 TWh; GDP, \$10.8 trillion.

continue to rise, with fossil fuels such as oil, natural gas, and coal playing the dominant role.

The intimate relationship between total energy use and economic output is further clarified in Fig. 3. One hopeful trend is that technological progress and efficiency improvements have reduced the energy intensity of economic production (i.e., lower requirements of physical energy per unit of economic output). Electricity will continue to play an increasingly important role as a safe, clean, and convenient form of energy.

3.1.2 Energy–Environment–Society Linkages

The environmental and social implications of energy use have not been as well analyzed as have energy–economy linkages. Although electricity has relatively few environmental and health consequences at the point of end use, key environmental and social issues arise from power generation. However, the extent and nature of impacts differ among energy sources. Oil- and coal-fired plants not only have national impacts but also have regional and global environmental and health effects. Complete life cycle analyses involving the mining, refining, processing, transport, conversion, and transformation of various fuels such as oil, coal, and nuclear materials all give rise to significant impacts. Even new and renewable energy sources such as wind power, geothermal energy, biomass, and ocean energy, which are perceived to be “clean,” have some negative social and environmental impacts.

3.1.3 Transnational Issues

Acid deposition is perhaps the most serious of the transnational issues faced today. Acid deposition is a result of oxides of sulfur and nitrogen that originate from fossil fuel combustion, falling to the ground as particulates and acid rain. Coal- and oil-fired power stations emit significant amounts of sulfur dioxide and nitrogen oxides into the atmosphere. The transport of sulfur dioxide occurs over distances of more than 1000 km, causing the deposition of emission products across national boundaries. Acid depositions caused by sulfur and nitrogen oxides result in damage to trees and crops and sometimes extend to acidification of streams and lakes, resulting in destruction of aquatic ecosystems. They also lead to the corrosion, erosion, and discoloration of buildings, monuments, and bridges. Indirect health effects are caused by the mobilization of heavy metals in acidified water and soil.

Other important transnational issues include environmental and health impacts of radiation due to severe nuclear accidents, oceanic and coastal pollution due to oil spills, downstream siltation of river water in one nation due to deforestation of water sheds and soil erosion in a neighboring country, and changes in hydrological flow and water conditions caused by dams.

3.1.4 Global Issues

The Intergovernmental Panel on Climate Change (IPCC) has identified energy use as the major

contributor to anthropogenic GHG emissions, mainly CO₂ and other gases such as N₂O, CH₄, and chlorofluorocarbons (CFCs) that will lead to climate change and undermine sustainable development prospects. First, global warming poses a significant potential threat to the future economic well-being of large numbers of humans. Second, climate change could also undermine social welfare and equity in an unprecedented manner. Third, from an environmental viewpoint, increasing anthropogenic emissions and accumulations of GHGs will significantly perturb a major global subsystem—the atmosphere. Climate change will also threaten the stability of a range of critical interlinked physical, ecological, and social systems and subsystems.

3.2 Framework for Sustainable Energy Development at the National Level

Sustainable development is the broad rationale underlying most national-level planning and policy-making. Power and energy planning must also be part of, and closely integrated with, overall sustainable development strategies to meet many specific, interrelated, and frequently conflicting national objectives. Some specific goals for sustainable energy development might include ensuring economic efficiency in energy supply and use to maximize growth, including energy efficiency-related objectives such as energy conservation; raising sufficient revenues from energy sales to finance sector development; addressing socioeconomic concerns such as meeting basic energy needs of the poor and developing special regions (particularly rural or remote areas) and priority sectors of the economy; preserving the environment; diversifying supply, reducing dependence on foreign sources, saving scarce foreign exchange, and meeting national security requirements; ensuring price stability; and so on.

3.2.1 Integrated Approach

Successful planning and implementation of national energy programs must explicitly link the energy sector to sustainable development of other parts of the economy. An integrated approach will help decision makers to formulate policies and will provide market signals and information to economic agents that encourage more efficient and sustainable energy production and use. Figure 4 depicts such an approach to decision making, within a hierarchical conceptual framework for sustainable energy development, to be implemented through a set of energy

supply and demand management policies. The middle column shows the core—an integrated, multilevel analytical framework. The top level of sustainable energy development recognizes transnational linkages. Thus, individual countries are embedded in an international matrix, and global economic, social, and environmental conditions will impose exogenous inputs or constraints on national decision makers.

The second hierarchical level in Fig. 4 focuses on the multisectoral national economy of which the energy sector is a part. Thus, energy planning requires analysis of links between the energy sector and other sectors, including energy needs of user sectors (e.g., industry, transport, agriculture), input requirements of the energy sector, and impacts of energy supply and pricing policies.

The next level of sustainable energy development disaggregates the energy sector into subsectors such as electricity and petroleum products. This permits detailed analysis, with special emphasis on interactions among various energy subsectors, substitution possibilities, and resolution of policy conflicts.

The lowest hierarchical level pertains to energy analysis within each energy subsector, where line institutions (both public and private) carry out detailed energy resource evaluation, planning, and implementation of projects (including sustainability assessments).

In practice, the various levels of sustainable energy development merge and overlap considerably, requiring careful study of (inter)sectoral linkages. Energy–social–environmental interactions (represented by the vertical bar in Fig. 4) cut across all levels and provide important paths for incorporating environmental and social considerations into national energy policies.

Sustainable energy development facilitates policy-making and does not imply rigid centralized planning. The process results in the development of a flexible and constantly updated sustainable energy strategy designed to meet national goals. This strategy (of which the investment program and pricing policy are important elements) may be implemented through energy supply and demand management policies and programs that make effective use of decentralized market forces and incentives.

In particular, SED implies improvements in overall economic efficiency through better energy management. Figure 4 shows various policy instruments available to decision makers for implementing sound energy management. While formulating policy, one

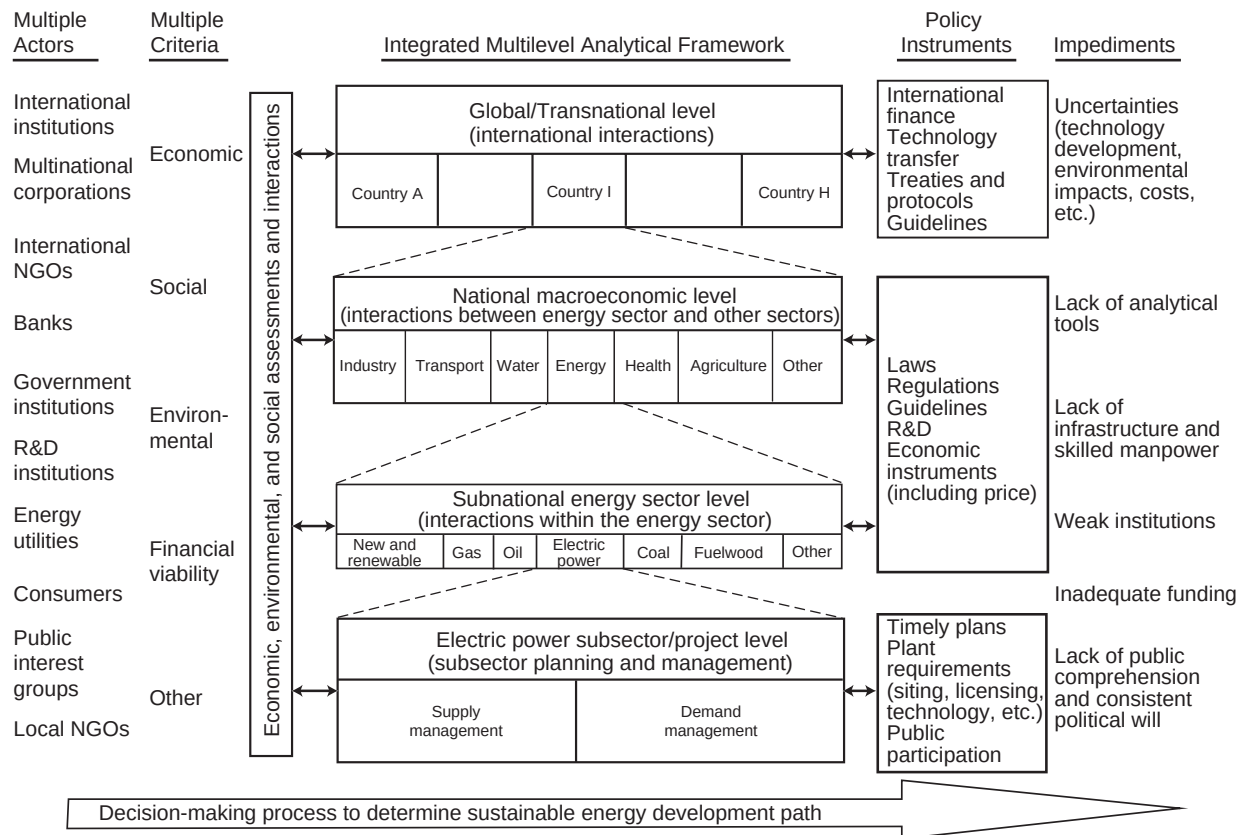


FIGURE 4 Framework for sustainable energy development. NGOs, nongovernmental organizations; R&D, research and development.

must consider the interests of multiple actors, ranging from international institutions to local energy users. The figure also indicates the most important impediments that limit the effectiveness of policies.

Investments offer a good opportunity to pursue sustainable energy development. In 10 years or so, new plants will account for more than half of the industrial output of developing countries, and in 20 years or so, new plants will account for practically all of it. Therefore, it will be possible to have a major impact by putting in place policies, legislation, mechanisms, systems, and incentives that facilitate sustainable energy development.

3.2.2 Identifying Sustainable Energy Options: "Win-Win" Options versus Trade-Offs

To identify sustainable energy options, policymakers need to consider the economic, social, and environmental aspects of sustainable development. Options that lead to improvements in all three indexes are referred to as "win-win" options. Once win-win

options are realized, policymakers are able to make trade-offs among other available options.

Incorporating environmental and social externalities into energy decision making is particularly important where concerns (e.g., pollution from nuclear or fossil-fueled plants, inundation at hydro plants) have hampered project implementation. Environmental and social concerns need to be addressed early, that is, at the sectoral and regional planning stages rather than at the final stage of project SDA. Unfortunately, when one deals with energy sector issues at this aggregate planning level, the application of many project-level valuation techniques becomes extremely difficult. First, the impacts are difficult to value (e.g., health effects of pollutants from coal-fired generating stations, biodiversity loss from large-scale hydro storage, impacts of GHG emissions). Doubts raised about the valuation techniques themselves divert attention away from critical policy trade-offs. Second, many techniques appropriate at the micro level are less effective at the sector level. Thus, contingent valuation is more valid where respondents can be asked specific

questions about local impacts of a particular project to which they can relate, and it is difficult to apply at the sector level where one deals with large numbers of technology, site, and mitigation options.

In countries where inappropriate policies have encouraged wasteful and unproductive uses of some forms of energy, better energy management could lead to improvements in economic efficiency (higher value of net output produced), energy efficiency (higher value of net output per unit of energy used), energy conservation (reduced absolute amount of energy used), and environmental and social protection (reduced energy-related environmental and social costs). However, it might not be possible to satisfy all of these goals simultaneously. For example, in some developing countries where existing levels of per capita income and energy consumption are very low, affordable energy might have a high priority to meet basic energy needs.

The economic efficiency criterion, which maximizes the value of net output from all scarce resources in the economy (including energy), normally subsumes purely energy-oriented objectives such as energy efficiency and conservation. Furthermore, costs arising from energy-related adverse environmental impacts may be included in the energy economics analytical framework to determine how many other benefits society should be willing to forgo so as to avoid environmental damage.

Energy use and production in developing countries can be improved in several ways to make them more sustainable. First, energy efficiency may be increased by supply- and demand-side improvements. Second, environmentally and socially more benign technologies can be introduced, including fuel switching and renewable energy sources. Finally, price, institutional, and regulatory reforms could contribute to sustainable energy development.

3.2.3 Sustainable Energy Development Options Matrix

Table I shows typical impacts of selected energy options on the three elements of sustainable development. Although efficient supply-side options have clear economic gains in terms of savings in capital investments and environmental benefits from reductions in GHG emissions that result from decreased energy supply, they do not have obvious social benefits. Efficient end use options, as shown in the case of an efficient fuelwood stove, have benefits relating to all three elements. Although advanced technologies, such as clean coal combustion technologies, help to reduce air pollutants (e.g., CO₂, NO_x)

TABLE I

Selected Sustainable Energy Development Options Matrix

Option	Impact		
	Economic	Environmental	Social
Supply efficiency	+	+	
End use efficiency	+	+	+
Advanced technologies	–	+	+
Renewables	–	+	+
Pricing policy	+	+	+/-
Privatization/decentralization	+	+/-	+/-

that cause respiratory diseases and reduce productivity, many developing countries cannot afford such high-cost technologies. Likewise, renewable energy sources also provide environmental and social benefits by reducing a country's dependence on traditional fossil fuels. However, in terms of power-generating costs, renewables are often more expensive than fossil fuels.

4. APPLYING THE SUSTAINABLE ENERGY DEVELOPMENT FRAMEWORK

This section presents practical case studies that illustrate the application of the ideas presented earlier.

4.1 Global Transnational Scale: Interplay of Optimality and Durability in Determining Global GHG Emission Target Levels

Optimization- and durability-based approaches can facilitate the determination of target GHG emission levels. Under an economic optimizing framework, the ideal solution would be, first, to estimate the long-run marginal abatement costs (MACs) and the marginal avoided damages (MADs) associated with various GHG emission profiles (see Fig. 5C where the error bars on the curves indicate measurement uncertainties). The optimal emission levels would be determined at the point where future benefits (in terms of climate change damage avoided by reducing one unit of GHG emissions) are just equal to the corresponding costs (of mitigation measures required to reduce that unit of GHG emissions), that is,

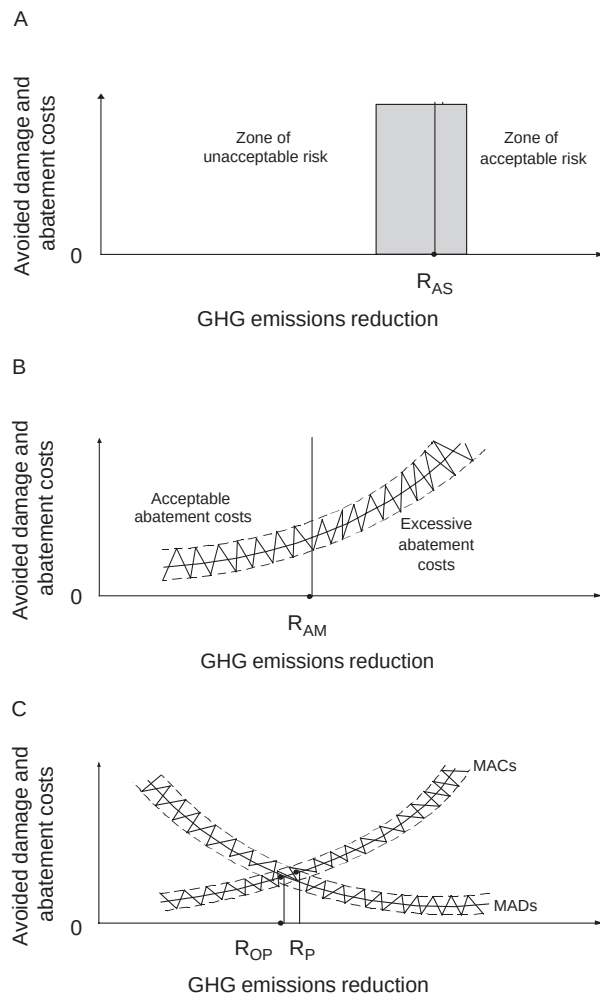


FIGURE 5 Determining abatement targets: (A) absolute standard; (B) affordable safe minimum standard; (C) cost-benefit optimum. Adapted from Intergovernmental Panel on Climate Change (1996, Fig. 5.10).

MACs = MADs at point R_{OP} . Durable strategies become more relevant when we recognize that MACs and/or MADs might be poorly quantified and uncertain. Figure 5B assumes that MACs are better defined than MADs. First, MACs are determined using technoeconomic least cost analysis, an optimizing approach. Next, the target emissions are set on the basis of the affordable safe minimum standard (at R_{AM}), which is the upper limit on costs that will still avoid unacceptable socioeconomic disruption. This is closer to the durability approach.

Finally, Fig. 5A indicates an even more uncertain world where neither MACs nor MADs are defined. Here, the emission target is established on the basis of an absolute standard (R_{AS}) or safe limit that would avoid an unacceptably high risk of damage to

ecological (and/or social) systems. This last approach would be more in line with the durability concept.

4.2 National Economy-wide Scale: Using the AIM for Macro Policy Analysis

National policymakers routinely make key macro-level decisions that could have (often inadvertent) environmental and social impacts that are far more significant than the effects of local economic activities. At this macroeconomic level, recent work has focused on incorporating environmental considerations such as depletion of natural resources and pollution damage into the system of national accounts via useful new indicators and measures such as the system of environmentally adjusted environmental accounts (SEEA), green gross national product (GNP), and genuine savings. In this context, the AIM approach is a sustainomics tool that helps to identify and analyze economic-environmental-social interactions and formulate effective sustainable development policies.

Table II shows a simplified AIM, although an actual one would be larger and more detailed. The AIM helps to promote an integrated view, meshing development decisions with priority economic, environmental, and social impacts. The left-most column lists examples of key development interventions (both policies and projects), whereas the top row indicates some typical sustainable development impacts/issues. Thus, the elements or cells in the matrix help to accomplish the following:

- Explicitly identify key linkages
- Focus attention on methods of analyzing important impacts
- Suggest action priorities and remedies

At the same time, the organization of the overall matrix facilitates the tracing of impacts as well as the coherent articulation of the links among a range of development actions, both policies and projects.

4.2.1 Screening and Problem Identification

One of the early objectives of the AIM-based process is to help in screening and problem identification by preparing a preliminary matrix that identifies broad relationships and provides a qualitative idea of the magnitudes of impacts. Thus, the preliminary AIM helps to prioritize the most important links between policies and their sustainability impacts.

One typical primary energy policy shown in the third row of Table II involves increasing energy

TABLE II

A Simplified Preliminary Action Impact Matrix

Activity or policy	Main objective	Impacts on key sustainable development issues			
		Land degradation and biodiversity loss	Air pollution GHG emissions	Resettlement and social effects	Other
Macroeconomic and sectoral policies	Macroeconomic and sectoral improvements	Positive impacts due to removal of distortions Negative impacts mainly due to remaining constraints			
Exchange rate	Improve trade balance and economic growth	(-H) (deforest open-access areas)			
Energy pricing	Improve energy use economic and efficiency		(+ M) (energy efficiency)		
Others					
Complementary measures	Socioeconomic and environmental gains	Enhance positive impacts and mitigate negative impacts (above) of broader macroeconomic and sectoral policies			
Market based			(+ M) (pollution tax)		
Non-market based		(+ H) (property rights)	(+ M) (public sector accountability)		
Investment projects	Improve effectiveness of investments	Investment decisions made more consistent with broader policy and institutional framework			
Project 1 (hydro dam)		(-H) (inundate forests)	(+ M) (displace fossil fuel use)	(-M) (displace people)	
Project 2 (Reafforest and relocate)		(+ H) (replant forests)		(+ M) (relocate people)	
Project N					

Note. Several typical policies and projects, as well as key economic, environmental and social issues, are shown. Some illustrative impact assessments are also indicated, where + and - signify beneficial and harmful impacts, and where H and M indicate high and moderate intensity, respectively.

prices closer to marginal costs to improve energy efficiency while decreasing air pollution and GHG emissions. A complementary environmental policy measure indicated in the fourth column consists of adding pollution taxes to marginal energy costs, a measure that will further reduce air pollution and GHG emissions. Increasing public sector accountability constitutes another complementary policy that will reinforce favorable responses to these price incentives by reducing the ability of inefficient firms to pass on cost increases to consumers or to transfer their losses to the government.

Near the bottom of Table II, a major hydroelectric project is shown as having two adverse impacts (inundation of forested areas and village dwellings) as well as one positive impact (replacement of thermal power generation, thereby reducing air

pollution and GHG emissions). A reforestation project coupled with resettlement schemes helps to address the negative impacts.

Therefore, this matrix-based approach encourages the systematic articulation and coordination of policies and projects to make development more sustainable.

4.2.2 Analysis and Remediation

This process may be developed further to assist in analysis and remediation. For example, more detailed analyses and modeling may be carried out for those matrix elements in the preliminary AIM that had already been identified as representing high-priority linkages between economy-wide policies and economic, environmental, and social impacts. This, in turn, would lead to a more refined and updated

AIM that would help to quantify impacts and formulate additional policy measures to enhance positive linkages and mitigate negative ones. The types of more detailed analyses that could help to determine the final matrix would depend on planning goals and available data and resources. They may range from fairly simple methods to rather sophisticated economic, ecological, and social models, included in the sustainomics toolkit.

4.3 Subnational Scale: Energy Sector Planning

The incorporation of environmental and social externalities into decision making is particularly important in the electric power sector. Meier and Munasinghe, in their Sri Lanka study, demonstrated how externalities could be incorporated into power system planning in a systematic manner. Sri Lanka currently depends largely on hydro power for electricity generation, but over the next decade or so the main choices seem to be large coal- or oil-fired stations or hydro plants whose economic returns and environmental impacts are increasingly unfavorable. In addition, a wide range of other options (e.g., wind power, increasing use of demand-side management, system efficiency improvements) complicates decision making, even in the absence of environmental concerns.

The methodology involves the following steps: (1) definition of generation options and their analysis using sophisticated least cost system planning models; (2) selection and definition of attributes that reflect planning objectives; (3) explicit economic valuation of those impacts for which valuation techniques can be applied with confidence, with the resultant values then being added to the system costs (the main economic attribute); (4) quantification of those attributes for which explicit economic valuation is inappropriate but for which suitable quantitative impact scales can be defined; (5) translation of attribute value levels into value functions (known as “scaling”); (6) display of trade-offs to facilitate decision making; and (7) definition of options for further study, also involving discarding eliminating inferior options.

4.3.1 Main Results

The main set of sectoral policy options examined included (1) variations in the currently available mix of hydro and thermal (coal and oil) plants, (2) demand-side management (e.g., compact fluorescent

lighting), (3) renewable energy options (e.g., wind generation), (4) improvements in system efficiency (using more ambitious targets for transmission and distribution losses than the base case assumption of 12% by 1997), (5) clean coal technology (e.g., pressurized fluidized bed combustion [PFBC] in a combined cycle mode), and (6) pollution control technology options (e.g., various fuel switching and pollution control options such as importing low-sulfur oil for diesels and fitting coal power plants with flue gas desulfurization [FGD] systems).

A limited number of criteria or attributes should be selected with care to reflect issues of national as well as local project level significance. CO₂ emissions were used as a proxy for the potential impact on global warming. Health impacts were measured through population-weighted increments in both fine particulates and NO_x. To capture the potential biodiversity impacts, a probabilistic index was derived. Employment creation was used as an illustrative social impact.

Figure 6A illustrates a typical trade-off curve for biodiversity. The “best” solutions lie closest to the origin. The so-called trade-off curve is defined by the set of “noninferior” solutions, representing options that are superior regardless of the weights assigned to the various objectives. For example, on this curve, the option defined as “no hydro” is better than the “wind” option in terms of both economic cost and biodiversity loss.

Although most options in Fig. 6A have index values that fall between 50 and 100, the no-hydro option has a zero value because thermal projects that constitute this option lie at sites of low biodiversity value (either close to load centers or on the coast). Meanwhile, wind plants would require large land areas, and their biodiversity indexes are higher. However, vegetation on the south coast (where wind power plants would be located) has relatively low biodiversity value; therefore, the overall biodiversity impact of this option is small. In summary, the best options (on the trade-off curve) include the no-hydro and run-of-river hydro options that require essentially zero inundation. The extreme outlier at the top right-hand corner is the Kukule hydro dam; it has a biodiversity loss index ($B = 530$) that is much larger than those for other options ($Bs = 50\text{--}70$).

The trade-off curve between health impacts and average incremental cost is different (Fig. 6B). The “iresid” point on this trade-off curve (low-sulfur imported fuel oil at diesel plants) is better than the use of flue gas desulfurization systems (“FGD” point) in terms of both economic cost and environment.

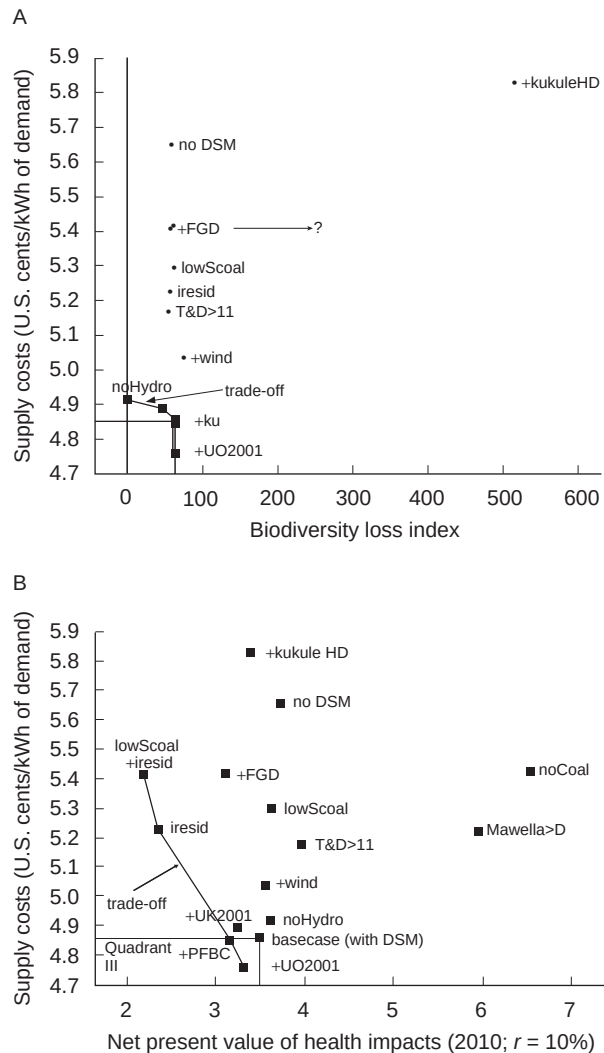


FIGURE 6 Trade-off curves between economic costs and (A) biodiversity impacts and (B) health impacts. Reproduced from Meier and Munasinghe (1994).

4.3.2 Conclusions

There are several useful conclusions. First, the results indicate that those impacts for which valuation techniques are relatively straightforward and well established (e.g., the opportunity costs of lost production from inundated land, benefits of establishing fisheries in reservoirs) are small compared with overall system costs, and so including them in the benefit–cost analysis does not materially change results. Second, even in cases where explicit valuation is difficult (e.g., mortality and morbidity effects of air pollution), implicit valuation based on analysis of trade-off curves can provide important guidance to decision makers. Third, certain options were clearly inferior or superior to other options when one examines all impacts simultaneously. For example, the high dam version of

the Kukule hydro project can be safely excluded from further consideration here due to poor performance on all attribute scales. Fourth, it is possible to derive attribute scales that provide useful proxies for impacts that are difficult to value. For example, the population-weighted, incremental, ambient air pollution level was the proxy for health impacts, yielding several important conclusions independent of any economic values assigned to health effects.

Finally, with respect to the practical planning, the study identified several priority recommendations, including the need to reconsider (1) demand-side management options, especially fluorescent lighting; (2) whether the current transmission and distribution loss reduction target of 12% ought to be further reduced; (3) possibilities of PFBC technology for coal power; (4) replacement of some coal-fired power plants (on the south coast) by diesel units; and (5) cooling system options for coal plants.

4.4 Local Project Scale: Multicriteria Analysis of Small Hydroelectric Power Projects

Well-accepted environmental and social assessment procedures at the project/local level may be readily adapted to assess environmental and social effects of micro-level activities. When monetary valuation of environmental and social effects is not feasible, MCA may be used. This subsection summarizes how Morimoto and colleagues have used MCA to compare hydroelectric power schemes. The three main sustainable development issues considered are the economic costs of power generation, the ecological costs of biodiversity loss, and the social costs of resettlement.

The principal objective is to generate additional kilowatt-hours of electricity to meet growing power demand in Sri Lanka. Assuming that the benefits from each additional kilowatt-hour are the same, the analysis seeks to minimize economic, social, and environmental costs of generating one unit of electricity from various hydropower sites. Following the MCA approach, environmental and social impacts are measured in different (nonmonetary) units instead of attempting to economically value and incorporate them within the monetary-valued cost–benefit analysis framework.

4.4.1 Environmental, Social, and Economic Indicators

Sri Lanka has many varieties of endemic or endangered fauna and flora. Often, large hydro projects destroy wildlife at dam sites and in

downstream areas. Hence, a biodiversity loss index was estimated for each hydroelectric site as the main ecological indicator.

Although dam sites are usually in less densely populated rural areas, resettlement is still a serious problem. In general, people are relocated from the wet zone to the dry zone where soils are less rich; therefore, the same level of agricultural productivity cannot be maintained. In the wet zone, multiple crops, including paddy, tobacco, coconuts, mangoes, onions, and chillies, can be grown. However, these crops cannot be cultivated as successfully in the dry zone due to limited water and poor soil quality. Living standards often become worse, and several problems (e.g., malnutrition) can occur. Moreover, other social issues might arise, including erosion of community cohesion and psychological distress due to changed living conditions. Hence, minimizing the number of people resettled due to dam construction is an important social objective.

The project costs are available for each site, from which the critical economic indicator (average cost per kilowatt-hour/year) may be estimated. The annual energy generation potential at various sites ranges from approximately 11 to 210 kWh (Fig. 7). All three variables (biodiversity index, number of people resettled, and generation costs) are weighted by the inverse of the amount of electrical energy generated. This scaling removes the influence of project size and makes them more comparable.

4.4.2 Basic Results

A simple statistical analysis shows that pairwise, there is a little correlation among the quantity of electricity generated, average generation cost, number of people resettled, and biodiversity index.

Figure 7 shows that on a per kilowatt-hour/year basis, the project named AGRA003 has the highest biodiversity index, HEEN009 has the highest

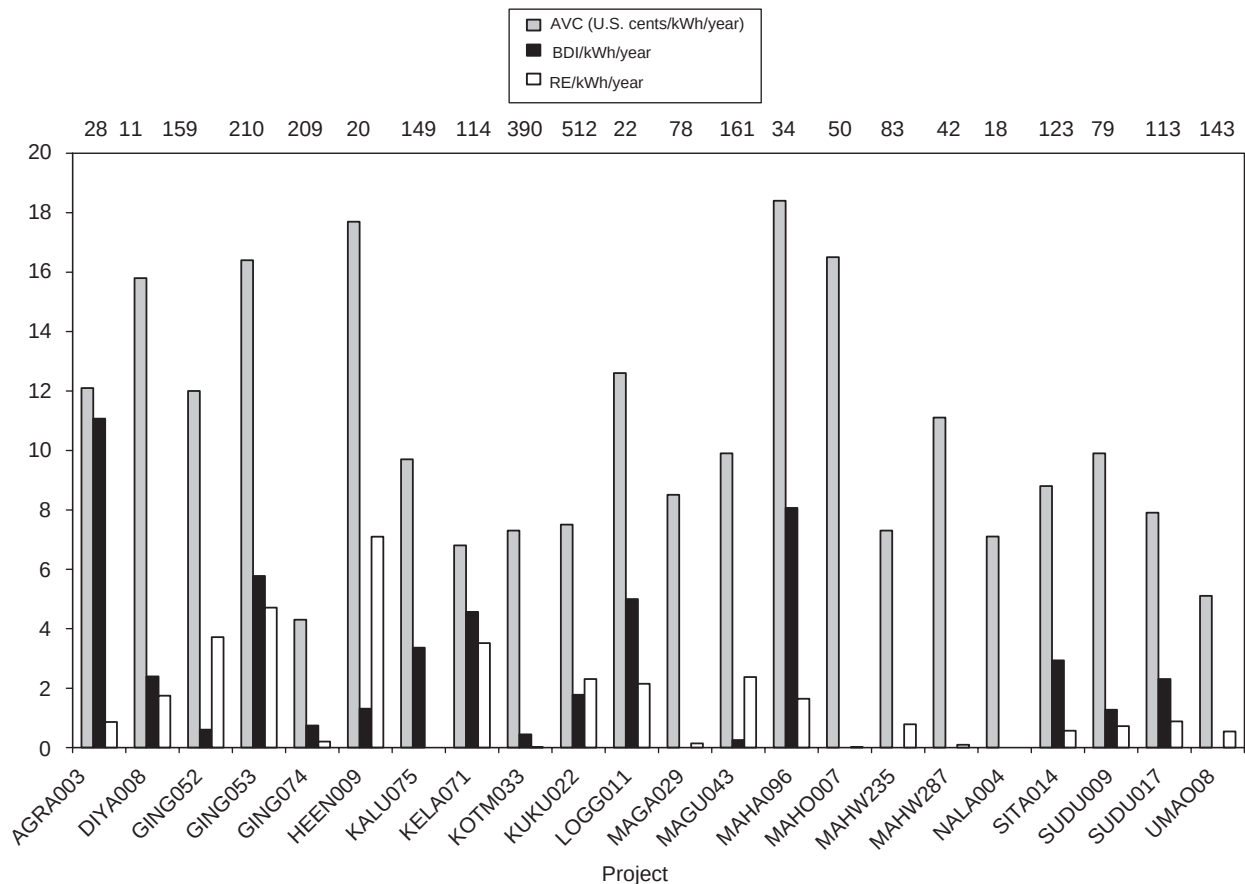


FIGURE 7 Average generation costs (AVC), biodiversity index (BDI), and number of resettled people (RE) by hydroelectric project. All indexes are per kilowatt-hour/year. Numbers of people resettled and the biodiversity index are scaled for convenience (by the multipliers 10^{-5} and 10^{-9} , respectively). The values across the top of the graph indicate the annual energy generation in gigawatt-hours. Adapted from Morimoto, Munasinghe, and Meier (2000).

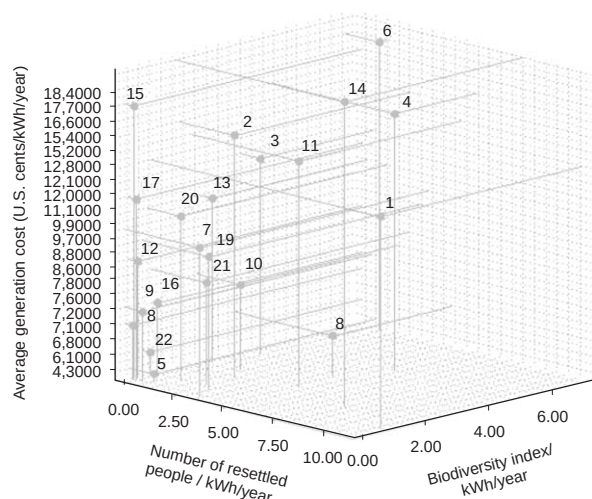


FIGURE 8 Three-dimensional MCA of sustainable development indicators for various hydropower options. Reproduced from Morimoto, Munasinghe, and Meier (2000).

number of resettled people, and MAHA096 has the highest average generation cost. Further important comparisons may be made. For example, KALU075 is a relatively large project whose costs are low, whereas MAHA096 is a smaller scheme with much higher costs with respect to all three indexes. Another observation is that a project such as KELA071 fully dominates GING053 given that the former is superior in terms of all three indicators. Similar comparisons may be made among other projects.

Figure 8 provides a three-dimensional analysis of sustainable development indicators for these hydropower sites where the respective axes represent economic, ecological, and social objectives. The closer to the origin any given coordinate point lies, the better the corresponding project is in terms of achieving these three objectives. This type of analysis gives policymakers some idea about which project is more favorable from a sustainable energy development perspective.

Suppose that all three objectives are arbitrarily given equal weight. Then, each project may be ranked according to its absolute distance from the origin. For example, rank 1 is given to the one closest to the origin, rank 2 is the second closest, and so on (Fig. 8). On this overall basis, from a sustainable energy development perspective, the most favorable project is GING074 (rank 1), whereas the least favorable is MAHA096 (rank 22).

4.4.3 Conclusions

The strength of this type of analysis is in helping policymakers to compare project alternatives more

easily and effectively. The simple graphical presentations are readily comprehensible and clearly identify sustainable development characteristics of each scheme. The multidimensional analysis supplements more conventional cost-benefit analysis (based on economic analysis alone). Because each project has different features, assessing them by looking at only one aspect (e.g., generation costs, effects on biodiversity, impacts on resettlement) could be misleading.

There are some weaknesses in the MCA approach used here. First, for simplicity, each major objective is represented by only one variable, assuming that all of the other attributes are minor. In reality, there may be several additional variables that could describe the economic, social, and environmental aspects of sustainable development. Further analysis that includes other attributes might provide new insights. A second extension of this study is to include other renewable sources of energy in the analysis. Finally, more sophisticated three-dimensional graphic techniques may yield better and clearer representations.

SEE ALSO THE FOLLOWING ARTICLES

Cultural Evolution and Energy • Development and Energy, Overview • Economic Growth and Energy • Economic Growth, Liberalization, and the Environment • Energy Ladder in Developing Nations • Green Accounting and Energy • Oil-Led Development: Social, Political, and Economic Consequences • Population Growth and Energy • Potential for Energy Efficiency: Developing Nations • Women and Energy: Issues in Developing Nations

Further Reading

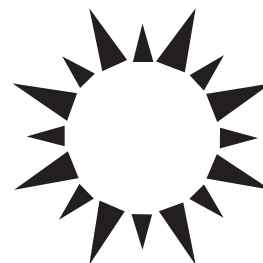
- Arrow, K. J., Cline, W., Maler, K. G., Munasinghe, M., and Stiglitz, J. (1995). Intertemporal equity, discounting, and economic efficiency. In "Global Climate Change: Economic and Policy Issues" (M. Munasinghe, Ed.). World Bank, Washington, DC.
- Atkinson, G., Dubourg, R., Hamilton, K., Munasinghe, M., Pearce, D. W., and Young, C. (1997). "Measuring Sustainable Development: Macroeconomics and the Environment." Edward Elgar, Cheltenham, UK.
- Commission on Sustainable Development. (1998). "Indicators of Sustainable Development." CSD, New York.
- Faucheux, S., Pearce, D., and Proops, J. (eds.). (1996). "Models of Sustainable Development." Edward Elgar, Cheltenham, UK.
- Gunderson, L., and Holling, C. S. (2001). "Panarchy: Understanding Transformations in Human and Natural Systems." Island Press, New York.

- IPCC. (1996). "Climate Change 1995: Economic and Social Dimensions," chap. 5. Cambridge University Press, London, UK.
- Meier, P., and Munasinghe, M. (1994). "Incorporating Environmental Concerns into Power Sector Decision Making." World Bank, Washington, DC.
- Morimoto, R., Munasinghe, M., and Meier, P. M. (2000). "Sustainable energy development: Assessing the Economic, Social, and Environmental Implications of Hydropower Projects in Sri Lanka. MIND Research-Discussion Paper NO.2." Munasinghe Institute for Development, Colombo, Sri Lanka.
- Munasinghe, M. (1992). "Environmental economics and sustainable development. Paper presented at the UN Earth Summit." Rio de Janeiro, Brazil, and reprinted by World Bank, Washington, DC.
- Munasinghe, M. (2002). The sustainomics trans-disciplinary meta-framework for making development more sustainable: Applications to Energy Issues. *Intl. J. Sustainable Dev.* 4(2), 6–54.
- Munasinghe, M., and Cruz, W. (1994). "Economywide Policies and the Environment." World Bank, Washington, DC.
- Munasinghe, M., Sunkel, O., and de Miguel, C. (Eds.). (2001). "The Sustainability of Long-Term Growth." Edward Elgar, London, UK.
- North, D. (1990). "Institutions, Institutional Change, and Economic Performance." Cambridge University Press, Cambridge, UK.
- Putnam, R. D. (1993). "Making Democracy Work: Civic Traditions in Modern Italy." Princeton University Press, Princeton, NJ.
- Rawls, J. A. (1971). "Theory of Justice." Harvard University Press, Cambridge, MA.
- Sen, A. K. (1984). "Resources, Values, and Development." Blackwell, Oxford, UK.
- United Nations. (1996). "Indicators of Sustainable Development: Framework and Methodology." UN, New York.
- UN Development Program. (1998). "Human Development Report." UN, New York.
- World Commission on Environment and Development. (1987). "Our Common Future." Oxford University Press, Oxford, UK.
- World Bank. (1998). "Environmental Assessment Operational Directive" (EAOD4.01). World Bank, Washington, DC.

System Dynamics and the Energy Industry

ANDREW FORD

Washington State University
Pullman, Washington, United States



-
1. System Dynamics
 2. Software
 3. Information Feedback
 4. Models for National Energy Policy
 5. Incentives for Electric Vehicles
 6. Power Plant Construction in the West
 7. Summary

Glossary

feebate Incentive program with fees on dirty cars and rebates for clean cars.

negative feedback A closed chain of cause and effect that acts to reduce change in a system.

positive feedback A closed chain of cause and effect that acts to magnify change in a system.

Stella Computer software for stock and flow modeling.

Vensim Computer software for stock and flow modeling.

widget A hypothetical product used in classroom examples.

System dynamics is a computer-based method for studying dynamic behavior. The method is normally applied when the dynamic behavior is viewed as a problem. The problematical behavior might take the form of highly volatile oscillations in prices, for example. System dynamics can help the analyst understand how the elements of the system interact to create the oscillations. The analyst would construct a mathematical model to represent the elements and their interactions. The model is then simulated on the computer to determine if simulated oscillations are similar to the oscillations observed in the industry. The analyst would then use the model to experiment with changes in the system that could lead to improved behavior. System dynamics has been applied to a wide variety of energy systems. The

applications have dealt with challenges faced by utility managers, state regulators, and national policy makers. This article presents a description of system dynamics followed by summaries of three applications to energy systems.

1. SYSTEM DYNAMICS

The system dynamics approach was pioneered by Jay W. Forrester and colleagues at the Sloan School of Management at the Massachusetts Institute of Technology during the 1960s. In 1977, Coyle classified system dynamics as “that branch of control theory which deals with socioeconomic systems and that branch of management science which deals with problems of controllability.” The method was well described by Forrester in 1961. In 2000, Sterman published the most comprehensive text.

The basic building blocks of system dynamics models are stocks and flows. Stocks represent where accumulation or storage takes place in the system. Flows represent the actions that directly influence the stocks. Figure 1 shows an example of a single stock to represent the volume of water stored in a reservoir. This stock is increased by the upstream flow and lowered by the flow of water released at the dam.

Figure 2 shows a combination of six stocks and six flows. The stocks represent the storage of petroleum products in the world oil industry. The flows show the movement of products from the oil fields of Kuwait to the fuel tanks in our personal vehicles.

Stocks and flows provide the foundation for system dynamics models. The next step is to add the variables needed to explain each flow. Figure 3 illustrates with a model of a sales company. A single stock is used to represent the number of salespersons, whereas flows are used to keep track of new hires

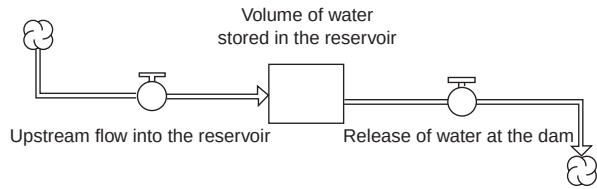


FIGURE 1 Example of a stock and flows. Adapted from A. Ford, "Modeling the Environment: An Introduction to System Dynamics Modeling of Environmental Systems." Copyright 1999 by Island Press. Reprinted by permission of Island Press.

and departures. The remaining variables are used to explain the flows. In this example, an exit rate is used to explain the departures, whereas a hiring fraction helps explain the new hires. New hires depend on the budgeted size of the sales force, which depends on the sales department budget and the average salesperson's salary. The sales department budget depends on the annual revenues and the fraction devoted to sales. Annual revenues depend on widgets sold and the widget price. Widget sales depends on the size of the sales force and the effectiveness of each salesperson.

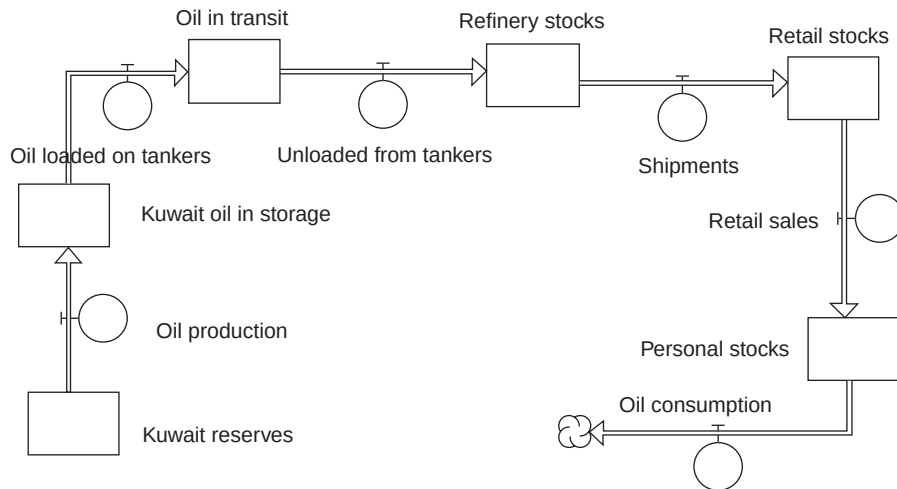


FIGURE 2 Example of stocks and flows in the petroleum industry. Adapted from Ford (1999). Reprinted by permission of Island Press.

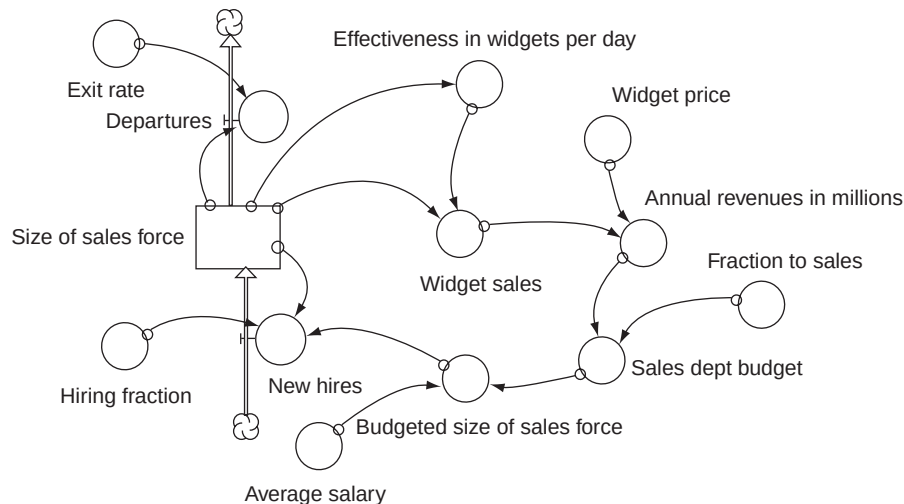


FIGURE 3 A system dynamics model of a hypothetical sales company. Adapted from Ford (1999). Reprinted by permission of Island Press.

In this classroom example, effectiveness is assumed to decline when there is a larger sales force.

2. SOFTWARE

The images in Figs. 1–3 were created with Stella software by High Performance Systems. Stella is one of several icon-based programs to aid the construction and simulation of system dynamics models. Stella diagrams are shown in many of the figures in this article, but equally useful models may be constructed with any of the stock-and-flow programs, such as Vensim or Powersim. The programs use somewhat different symbols, as shown by the Vensim diagram of the sales model in Fig. 4. Although the symbols may differ, the programs are fundamentally the same. They are all designed to facilitate the construction of stock-and-flow models and to show dynamic behavior through numerical simulation.

3. INFORMATION FEEDBACK

The sales company will grow rapidly if the personnel are highly effective at selling widgets. The rapid growth is made possible by positive feedback, a closed chain of cause and effect that acts to magnify change in a system. In this example, an increase in the sales force can lead to higher sales, higher revenues, increased budgets for the sales department,

more new hires, and an even larger sales force in the future. Understanding positive feedback is crucial if we are to understand the potential for systems to grow in a rapid manner.

However, as the company grows larger, each person is assumed to be less effective at selling widgets. This assumption introduces negative feedback, a closed chain of cause and effect that acts to reduce the change in a system. In this example, an increase in the sales force can lead to lower effectiveness, lower sales, less revenues, lower budgets for the sales department, fewer new hires, and a smaller sales force in the future. Understanding negative feedback is crucial if we are to understand the limits and stability of systems.

Figure 5 shows the combination of positive and negative feedback loops at work in the sales company. This diagram uses words for each of the variables and arrows for the causal connections. The arrows are arranged to allow the eye to easily follow the closed chains of cause and effect. A plus sign in the middle of the loop denotes the positive feedback that allows the company to grow. The minus signs in the middle of the loops denote the three negative loops in the sales model.

System dynamics is based on the premise that understanding dynamic behavior requires one to understand the interplay between the feedback loops in a system. A closely related tenet is that our mental skills are not sufficient to interpret the feedback structure of systems. It appears that we are good at

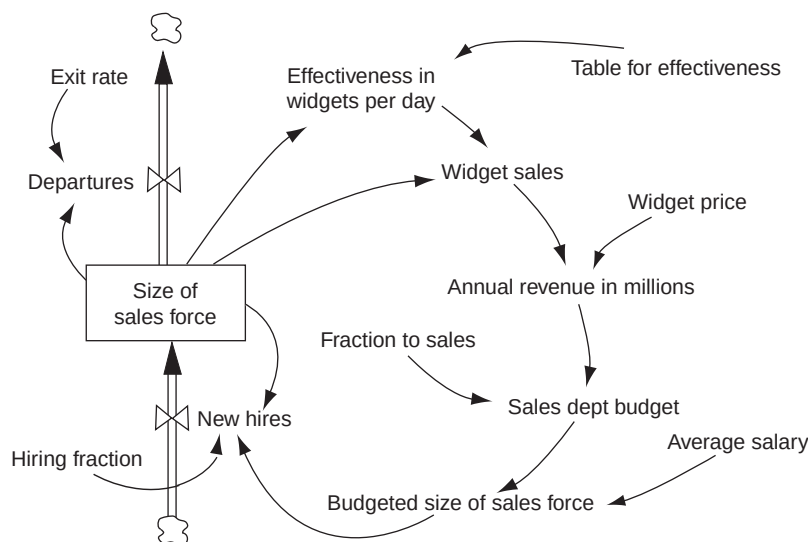


FIGURE 4 Vensim diagram of the sales company model. Adapted from Ford (1999). Reprinted by permission of Island Press.

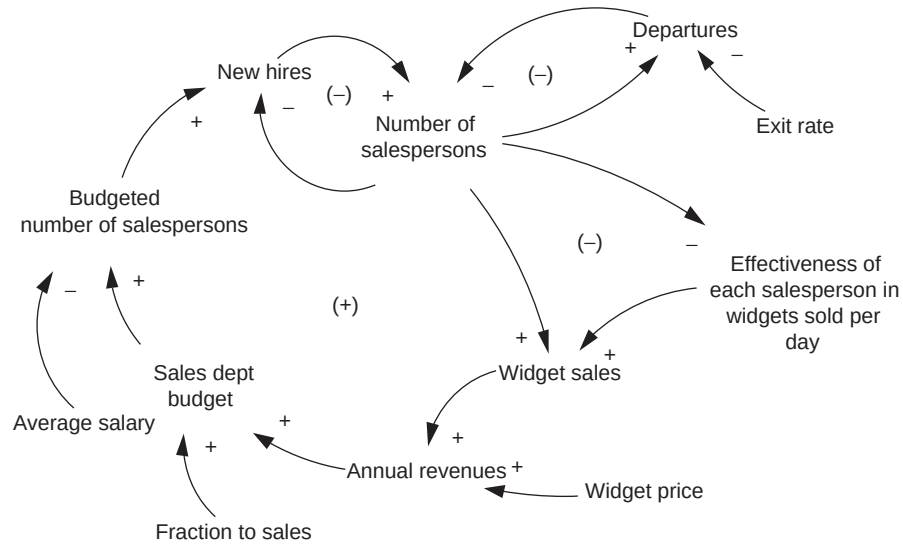


FIGURE 5 Feedback loops in the model of the sales company. Adapted from Ford (1999). Reprinted by permission of Island Press.

making lists of the many factors that might be important in a system, but we are limited in our ability to simulate the interaction between the key factors. Also, it appears that our mental skills are not sufficient to anticipate how the interactions in a system lead to growth, decay, oscillations, or other patterns of dynamic behavior.

System dynamics models are constructed with the goal of improving our understanding of dynamic behavior and anticipating the impact of policies that could lead to improved behavior. This goal has been reached in a wide range of energy models, including those that deal with privatization of electricity in the United Kingdom and deregulation of electricity in the United States. Other applications deal with variability of electricity supply in Columbia and conservation policy in the northwest region of the United States. In the following sections, three applications are described to convey a sense of the usefulness of the approach.

4. MODELS FOR NATIONAL ENERGY POLICY

One of the earliest applications to energy systems was Roger Naill's 1973 model of the exploration and production of natural gas in the United States. The model assigned one stock to represent the proven reserves of natural gas and a second stock for the unproven reserves. The model simulated the natural gas discoveries and production from 1900 to 2020 with different assumptions about the size of the

resource base and different policies on price regulation. The model was used to illustrate how the resource could be depleted in a surprisingly short interval when demand is growing in an exponential fashion.

The natural gas model was expanded to represent all of the fossil fuels—oil, coal, and natural gas. The expanded model simulated end-use demands for fossil fuels as well as the intermediate demands in the electricity sector. The new model, known as FOSSIL1, was much larger than the previous model, so space does not allow for a description of the many stocks and flows. The model focused on the difficult transition from conventional fossil fuels (e.g., oil) to ultimate sources (e.g., renewables). The most important application of FOSSIL1 was a simulation analysis of the policies advocated in the first National Energy Plan issued by President Carter in April 1977. The simulations indicated that the plan would “fall short of solving the nation’s energy dilemma,” but it would delay world oil shortages by approximately 10 years.

Naill and colleagues were asked to expand the model for policy support at the Office of Analytical Services of the U.S. Department of Energy. The expanded model was called FOSSIL2. The new model retained the focus on the difficult transition from conventional to ultimate sources of energy, but it was greatly expanded to represent the many fuels, technologies, regulations, and policies of interest to the Department of Energy. A typical simulation covered the interval from 1900 to 2100. In 1992, Naill described the model’s development and applications during both the Carter and the Reagan

administrations. He explained that system dynamics proved to be well suited for the analysis of the national energy system because of the ease of simulating stocks and flows and the importance of representing information feedback acting primarily through energy prices.

5. INCENTIVES FOR ELECTRIC VEHICLES

Another application focused on electric cars and their impact on an electric utility in California in the 1990s. Utility companies were considering incentives to promote the sale and use of electric vehicles. This is an intriguing idea but not a new one. It was first proposed by Thomas Edison, who encouraged the power companies of his day to build garages for electric vehicle owners. In the 1990s, California utilities proposed giving their customers financial incentives to reduce the purchase price of an electric vehicle. Increased sales of electric vehicles would lead to higher electricity consumption during nighttime hours when the demand could be satisfied by idle generating capacity. Utilities argued that incentives could lead to cleaner air and possibly lower electric rates. The reduction in electric rates might arise if the utility earned more in revenues from electricity sales than it would spend on fuel for the extra electricity generation. A system dynamics model was constructed to represent the vehicle purchase decisions of drivers and the accumulation of electric vehicles in the utility company's service territory. The model also simulated the operation of the company's power generators and the cash flows associated with financing the incentive program. The simulations suggested that it would be extremely difficult for the utility company to finance the incentives without having to raise its retail rates. The study concluded with the suggestion that feebates might be a more suitable method of promoting the sale of electric vehicles.

A feebate program would be run by a state or air shed regulator. The "fee" in feebates represents a fee imposed on the purchase price of a dirty vehicle. The "bate" represents a rebate to the person who purchases a clean vehicle. The idea is to finance the rebates for the clean vehicles with the fees imposed on the dirty vehicles. Proponents argue that feebates could be designed to reward cleaner vehicles, regardless of their technology or fuel. If one were thoughtful about the relative size of the fee and rebate, feebates might be designed with the cash

inflow from fees balancing the cash outflow for rebates.

System dynamics was used to study the dynamic challenges faced by a program manager who would be charged with promoting the sale of cleaner vehicles while keeping the cash flows in the program fund in approximate balance. The feebate model was adapted from the model of the California utility company and designed for ease of use by students as well as industry experts. The model is sufficiently compact that the key stocks and flows may be shown here.

Figure 6 shows vehicle sales and retirements along with annual travel and annual emissions. A stock of cars is used to represent the cars in use in a hypothetical air shed. The model distinguishes between cars fueled by gasoline, methanol, electricity, and compressed natural gas, with market shares estimated with a discrete choice model based on stated preferences in a survey of Californians. The Stella diagram in Fig. 6 shows three-dimensional icons to represent variables with six different values, one for each of the six types of vehicles in the model.

Figure 7 shows the portion of the model simulating the cash accumulating in the state fund. The main cash flow into the fund is the collection of fees imposed on people who purchase dirty vehicles. The main cash outflow is rebates paid to people who elect to purchase cleaner vehicles. As in Fig. 6, three-dimensional icons are used to represent variables with six different values, one for each type of vehicle.

The feebate model has been designed as a flight simulator, a highly interactive version of a system dynamics model to promote learning through interactive experimentation. Flight simulators have proven useful in a variety of business systems, including health care, airlines, insurance, real estate, and project management. The feebate flight simulation has proven useful primarily in the classroom, where it helps students experiment with their ideas of how to control a feebate program. Figure 8 shows the flight simulator part way through a simulation in which the student has set no fees and no rebates. This "do-nothing" simulation will provide a reference case for comparison with subsequent simulations with an active feebate programs.

The students' challenge is to introduce fees and rebates to reduce emissions while keeping the balance in the state fund within reasonable limits. This is a controllability challenge, one well suited for that "branch of management science which deals with problems of controllability."

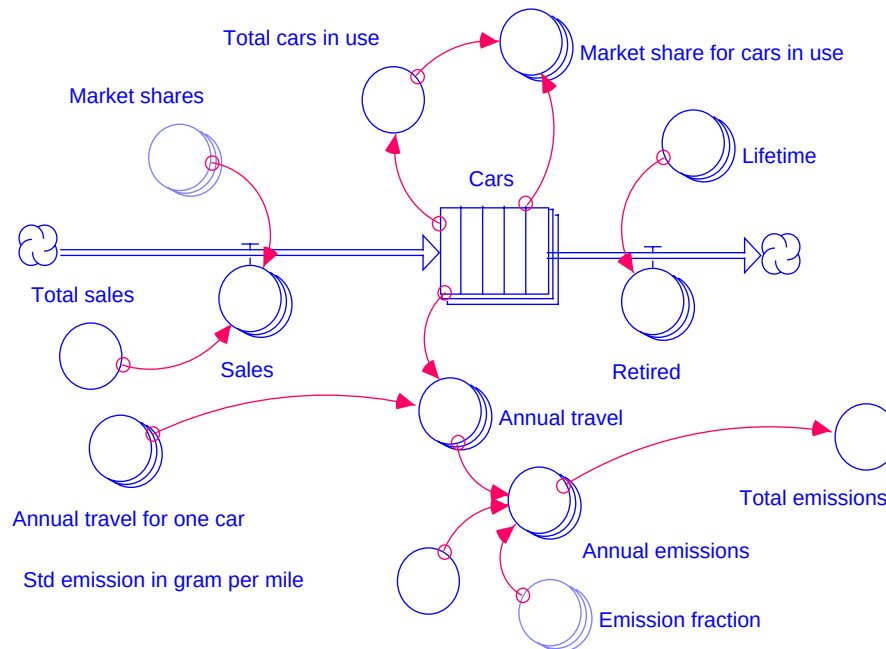


FIGURE 6 Simulating the sales and use of cars in the feebate model. Adapted from Ford (1999). Reprinted by permission of Island Press.

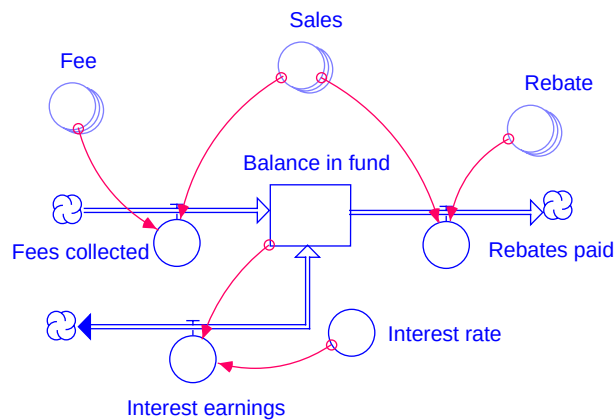


FIGURE 7 Simulating cash flows in the feebate model. Adapted from Ford (1999). Reprinted by permission of Island Press.

6. POWER PLANT CONSTRUCTION IN THE WEST

The third application deals with power plant construction in the deregulated electricity markets and is described in a special issue of the *Journal of Industry and Trade* devoted to the California electricity crisis. Many believe that competitive electricity markets are prone to the same cycles of boom and bust that

appear in commodity markets and in a specialized industry such as real estate. The system dynamics study demonstrated how boom and bust might appear in the western electricity system. The model simulates wholesale electricity market prices in the western United States, as explained in Fig. 9.

Figure 10 shows the simulated construction in a business-as-usual scenario. The model represents the construction of gas-fired combined cycle (CC) power plants by merchant investors. The curves in Fig. 10 are simulation results; the small rectangles are information buttons positioned to represent actual investments in CCs. The construction curve in Fig. 10 shows that the number of CCs under construction peaks near the end of 2001. From this point forward, construction completions are greater than construction starts, and the total megawatts of CCs under construction declines. This simulation suggests that we were at the crest of a building boom in 2001. As the construction is completed during 2002 and 2003, installed CC capacity will increase to approximately 27,000 MW by the start of 2004. Figure 10 shows a second wave of construction beginning in 2006 and peaking near the end of 2007.

This simulation scenario, created in November 2001, suggests that we were in the midst of a building boom, with the actual construction well

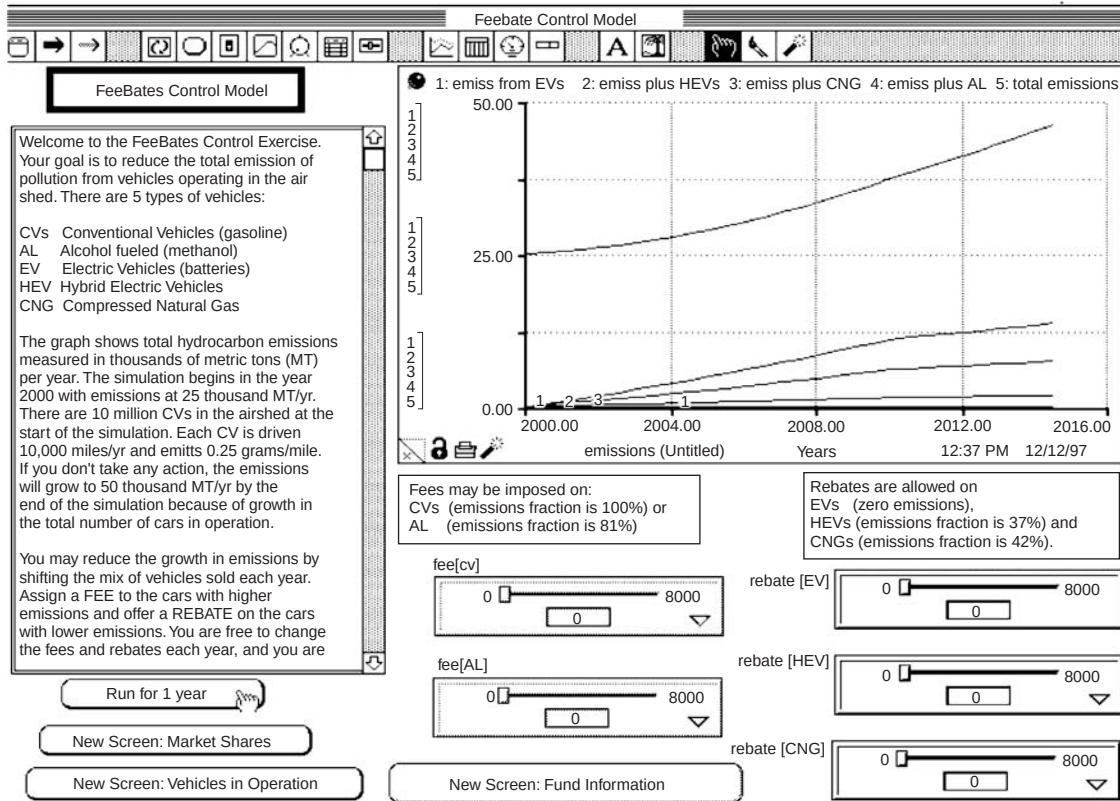


FIGURE 8 Results from a “do-nothing” simulation with the feebate model. Adapted from Ford (1999). Reprinted by permission of Island Press.

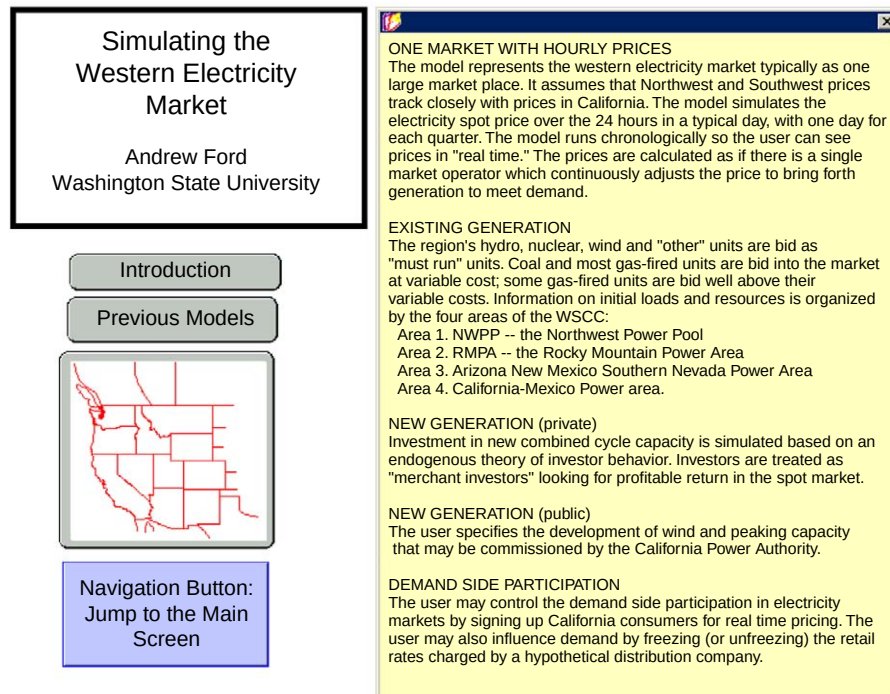


FIGURE 9 Opening screen of the western market model.

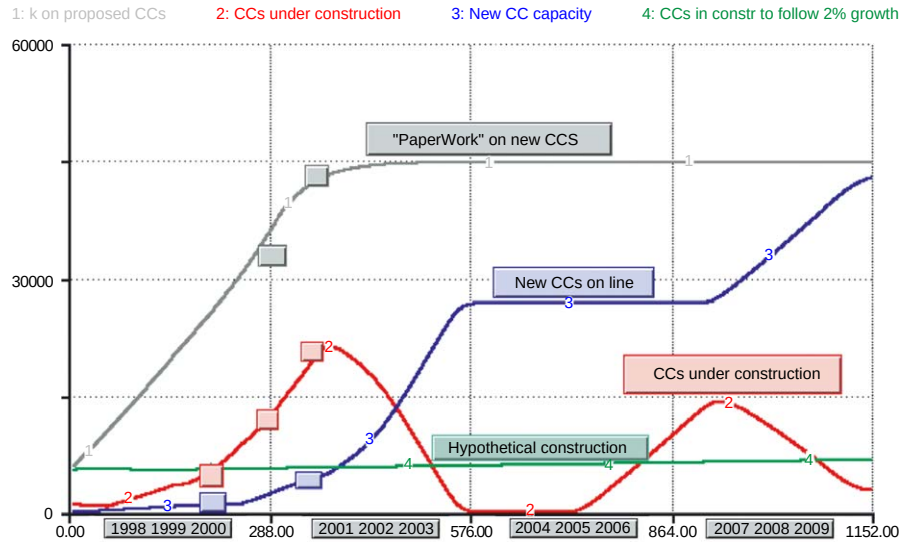


FIGURE 10 Simulated construction in the business-as-usual scenario.

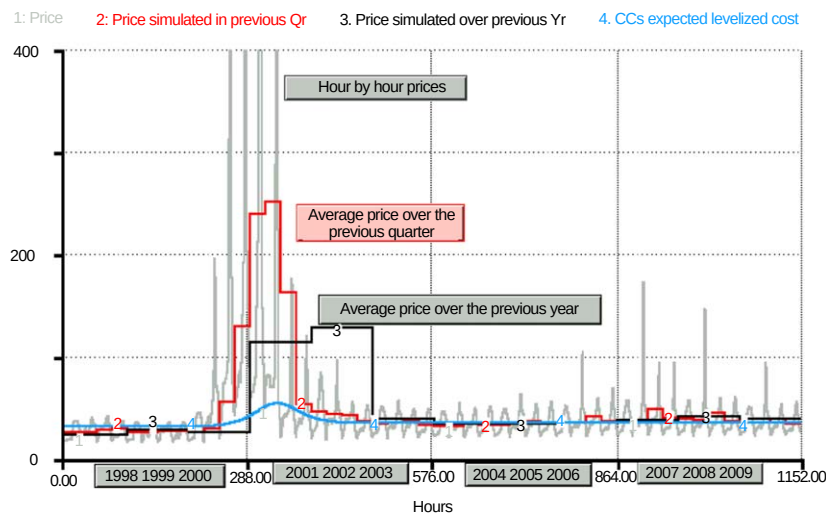


FIGURE 11 Wholesale market prices in the business-as-usual scenario.

above the hypothetical construction. This overbuilding situation continues until midway through 2003. For the next 3 years, however, the scenario envisions construction below the levels needed to keep pace with demand. During this interval, investors are reluctant to start construction, even though they hold a huge number of approved permits. Investors hold off on construction starts until 2006–2007.

Figure 11 shows market prices in the business-as-usual scenario, with the vertical scale running from \$0 to 400/MWh. The simulation indicates that hourly variations for typical days in 2002 and 2003 would be much smaller than the variations in 2001.

Figure 11 shows that quarterly prices continue to decline during the interval from 2002 to 2004. The construction boom allows for much lower prices for the interval from 2002 to 2005. The figure shows that price spikes could reappear as soon as 2006 and 2007. The spikes are much less severe than the spikes in 2000 and 2001. The improved behavior arises because the scenario is based on average gas and hydro conditions. For example, gas prices do not skyrocket in 2006 like they did in 2000. Also, hydro generation does not decline in 2006 and 2007, like it did in 2000 and 2001. These assumptions mask the underlying deterioration of the

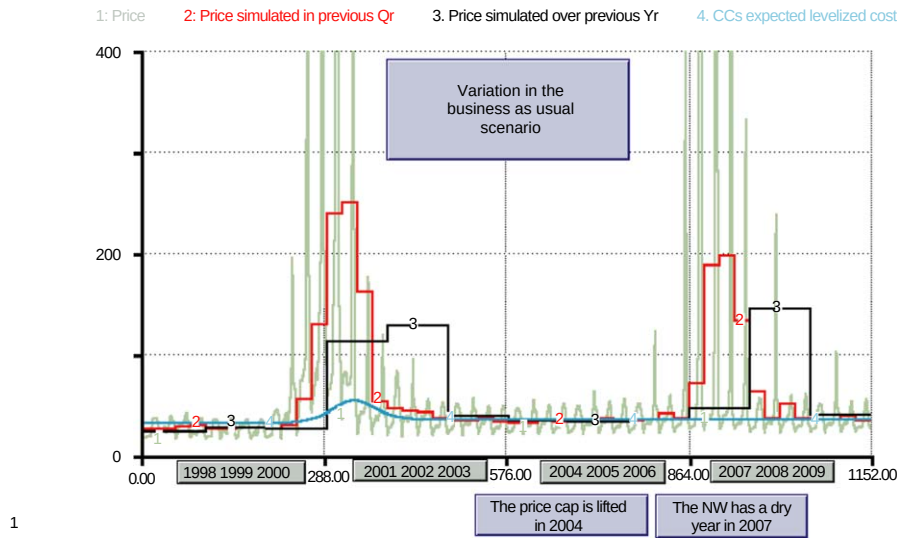


FIGURE 12 Market prices in a variation of the business-as-usual scenario.

supply–demand balance that occurs during the lull in construction.

Figure 12 shows a variation in the opening scenario to unmask the difficult conditions that could emerge within the next decade. The new simulation assumes that the northwest experiences a dry year in 2007 and that the price cap is lifted at the start of 2004. (We might envision the cap would be eliminated because those who oppose price caps would win adherents to their view during the previous 2 years—years of low prices and declining construction.) Figure 12 shows hourly prices spiking off the chart in 2007. Quarterly prices would climb to approximately \$200/MWh, and the average annual price for 2007 would increase to \$146/MWh, a value that exceeds the annual prices seen in either 2000 or 2001. Reserve margins (not shown here) would decline to levels lower than the simulated reserves in 2000 and 2001.

These simulations were used to promote discussion of fundamental changes in wholesale electricity markets. Without fundamental changes, the next construction boom would occur too late to prevent a decline in reserve margins and the reappearance of price spikes. If we continue with the current market structure, we run the risk of exposing the western electricity markets to another round of reliability alerts and price spikes. Alternative scenarios were generated to show the benefits of capacity payments to encourage private investment and state funding to finance public investment. On the demand side, simulation scenarios showed the

benefits of implementing real-time pricing in California.

7. SUMMARY

System dynamics has been applied to a variety of energy systems at the local, state, and national level. Previous applications indicate that the method is valued for the clarity of stocks and flows and the ease of simulating the feedback loops that give rise to dynamic behavior.

SEE ALSO THE FOLLOWING ARTICLES

Complex Systems and Energy • Computer Modeling of Renewable Power Systems • National Energy Modeling Systems • Neural Network Modeling of Energy Systems

Further Reading

- Coyle, G. (1977). "Management System Dynamics." Wiley, New York.
- Ford, A. (1994). Electric vehicles and the electric utility company. *Energy Policy* 22(7), 555–570.
- Ford, A. (1997). System dynamics and the electric power industry. *System Dynamics Rev.* 13(1), 57–85.
- Ford, A. (1999). "Modeling the Environment: An Introduction to System Dynamics Modeling of Environmental Systems." Island Press, Washington, DC.

- Ford, A. (2001). Simulation scenarios for the western electricity market, discussion paper for the California Energy Commission workshop on alternative market structures for California. Available at <http://www.wsu.edu/~forda>.
- Ford, A. (2002). Boom & bust in power plant construction: Lessons from the California electricity crisis. *J. Industry Competition Trade* 2(1/2), 59–74.
- Forrester, J. (1961). “Industrial Dynamics.” Pegasus, Waltham, MA.
- Morecroft, J., and Sterman, J. (1994). “Modeling for Learning Organizations.” Pegasus, Waltham, MA.
- Naill, R. (1992). A system dynamics model for national energy policy planning. *System Dynamics Rev.* 8(1), 1–19.
- Naill, R., and Backus, G. (1977). Evaluating the national energy plan. *Technol. Rev.*, July/August.
- Sterman, J. (2000). “Business Dynamics.” Irwin McGraw-Hill, New York.