MAGLEV Technology Development

James R. Powell[1], Gordon Danby[2]
[1]Maglev-2000 Corporation, Shoreham, NY, USA
[2]Maglev-2000 Corporation, Wading River, NY, USA

Article Outline

Glossary

Gap The physical clearance between a Maglev vehicle and the guideway it travels along.

Guideway The structure that holds Maglev vehicles as they travel. Guideways can be elevated above the local terrain, or on-grade, depending on the design, need for isolation from the local surroundings, and other factors.

Maglev (magnetic levitation) Transport of vehicles that are magnetically levitated above a guideway and magnetically propelled along it.

Magnetic Induction The generation of induced current in a conducting loop by the movement of a magnet past it.

Magnets, electromagnetic The generation of a magnetic field by a current in a loop of conductor, usually wound around an iron core to increase the strength of the magnetic field.

Magnets, permanent A physical object that is the source of a magnetic field, without the need for an electric current to produce the field. The movement of orbiting electrons in the atoms of the permanent magnet is locked in place to create the magnetic field.

Magnets, superconducting The generation of powerful magnetic fields from a loop or loops of zero electrical resistance superconducting wire, usually wound as an air core magnet without iron.

Superconductors Materials that when cooled below a transition temperature completely lose electrical resistance. They can carry very large currents at very high current densities, hundreds of thousands of amps per square centimeter of cross section, indefinitely without any decay in their current and without any voltage being applied to them.

Definition of the Subject and Its Importance

Maglev (magnetic levitation) is the first new mode of Earth-based transport since the invention of the airplane. Maglev vehicles are magnetically levitated and propelled along a guideway, with no mechanical contact and friction, their speed limited only by atmospheric drag. In evacuated tunnels, Maglev vehicles can potentially reach orbital speeds of 8 km/s (18,000 mph). Traveling in the atmosphere, Maglev vehicles have reached 580 km/h (361 mph).

Maglev vehicles have no engines, do not burn oil or other fuels, do not emit pollutants and greenhouse gases, and are quiet, with no rail or engine noise. They are propelled by the interaction between AC currents in the guideway and the magnets on the vehicle. The speed and position of Maglev vehicles on the guideway is determined by the frequency of the AC current that propels them. The distance between individual Maglev vehicles on the guideway can be designed to always remain constant, regardless of what conditions the different vehicles encounter, that is, up or down grades, curved guideways or straight, head or tail winds, etc. The individual vehicles are not

Robert A. Meyers (ed.), *Encyclopedia of Sustainability Science and Technology*, DOI 10.1007/978-1-4419-0851-3,

controlled by on-board engineers. Instead a central traffic control center views the entire guideway in real time to detect potential hazards, controlling the location and speed of all vehicles by means of the frequency of the AC propulsion current fed to the energized block of guideway that each individual vehicle travels on.

Maglev vehicles can travel as individual single units on the guideway, or during high levels of traffic, can be coupled together to form multi-vehicle consists. Each vehicle is individually propelled by the AC current applied to the energized block of guideway the vehicle, or the multi-vehicle, consist is traveling on.

There are two types of Maglev: electromagnetic and electrodynamic. In electromagnetic Maglev the vehicles are equipped with conventional electromagnets in which a DC current winding energizes an iron structure to create a strong magnetic field. The vehicle's electromagnets are positioned underneath two iron rails attached to the guideway. The magnetic force between the vehicle's electromagnets and the guideways iron rails located above them levitate the vehicle.

Electromagnetic Maglev has three major limitations. First, its magnetic suspension is inherently unstable. As the vehicle electromagnets move closer to the iron rails above them, the attractive force automatically increases. To prevent a crash, the DC current in the vehicle's electromagnets must be reduced by a servo control system that operates on a very fast time scale, that is, thousandths of a second. The servo control system thus continuously adjusts current to maintain a safe gap between the vehicle and the guideway. The second limitation is that to keep the DC magnet power required to levitate the vehicle at an acceptable level, the gap between the vehicle and guideway must be very small, on the order of 1 cm (3/8 in.). This requires a very precise and high cost guideway. The third limitation is that the lifting capability of the Maglev vehicle is small, because of the heavy magnet and power system weights.

The second type, electrodynamic Maglev, uses permanent magnets on the vehicle, not electromagnets. Instead of having iron rails on the guideway, electrodynamic Maglev uses ordinary aluminum loops that are located beneath the vehicle's permanent magnets, not above them. As the Maglev vehicle moves along the guideway, its magnets induce currents in the aluminum loops. These induced currents magnetically interact with the vehicle's permanent magnets, automatically levitating it. The suspension is automatically and inherently stable – the closer the vehicle gets to the aluminum loops beneath it on the guideway, the stronger the levitating force. With appropriate design, Maglev vehicles using electrodynamic forces can be strongly stable in the vertical, horizontal, pitch, yaw, and roll directions, but able to move freely in the longitudinal direction along the guideway.

Electrodynamic Maglev can use either conventional permanent magnets or superconducting magnets. The conventional permanent magnets are much weaker in magnetic strength than superconducting magnets. As a consequence, Maglev vehicles that use permanent magnets have very small gaps between the vehicle and the guideway, on the order of 1 cm (3/8 in.) and are very limited in their capability to lift heavy weights.

In contrast, Maglev vehicles that use superconducting magnets have much larger gaps between the vehicle and the guideway, on the order of 10 cm (4 in.) or more, and have much greater weight-lifting capability. Moreover, they are inherently very strong, stable, and can be designed to resist hurricane force winds and strong earthquakes without contacting the guideway.

The technology for superconducting magnets is commercially well developed. There are thousands of superconducting MRI medical units all around the world, and superconducting magnets are widely used in R&D laboratories. The Large Hadron Collider (LHC) at the CERN Laboratory in Switzerland, for example, has 54 km (31 miles) of superconducting magnets, all of which must function perfectly for the LHC to operate. The same length of magnets on Maglev vehicles would provide half of the highway and airplane passenger miles in the USA, plus half of the intercity truck long distance miles.

Japan is now operating a first-generation passenger Maglev system in Yamanashi Prefecture that has carried well over 80,000 passengers, with accumulated running distances of hundreds of thousands of kilometers. Japan plans to extend their system to become a 500 km (300 miles) route between Tokyo and Osaka that will carry 100,000 passengers daily, with a trip time of 1 h.

Germany has developed a first-generation electromagnetic Maglev passenger system called Transrapid.

A commercial 31 km Transrapid system now operates between the center of Shanghai, China, and its airport.

An advanced second-generation superconducting Maglev system, termed Maglev-2000, is being developed in the USA. The basic components – quadrupole magnets, aluminum loop guideway panels, guideway beam and vehicle – have been fabricated and successfully tested, but as yet funds for operational testing at high speed on a guideway have not been available.

The second-generation Maglev-2000 system will have important new and unique capabilities beyond those already achieved by the present Japanese and German first-generation Maglev systems that will make it a major mode of sustainable transport in the rest of the twenty-first century. These new capabilities include:

- Transport of autos and trucks on Maglev vehicles, enabling electric vehicles to make long trips without the need to consume gasoline
- Much lower cost guideway system that can be privately financed and will not require government subsidies
- High speed electronic skip-stop switching of vehicles to off-line stations for unloading/loading operations enabling high average speeds with many closely spaced stations
- Ability to adapt Maglev-levitated travel to existing railroad tracks at very low cost, enabling Maglev travel inside urban and suburban areas without needing to build new infrastructure
- Ability to greatly reduce highway deaths and injuries by transporting people, autos, and trucks on Maglev rather than on highways
- Ability to use superconducting Maglev for other important applications, including low cost storage of electrical energy from variable wind and solar farms; the long distance, low cost transport of fresh water to drought areas; and very low cost for launching payloads into orbit, particularly space solar power satellites that can beam massive amounts of clean power down to Earth to meet power needs in an environmentally acceptable manner

The implementation of Maglev as a major mode of transport will enable much better sustainable living standards for the world's population, without environmentally damaging the planet by the massive use of fossil fuels.

Introduction

History and Impact of New Modes of Transportation

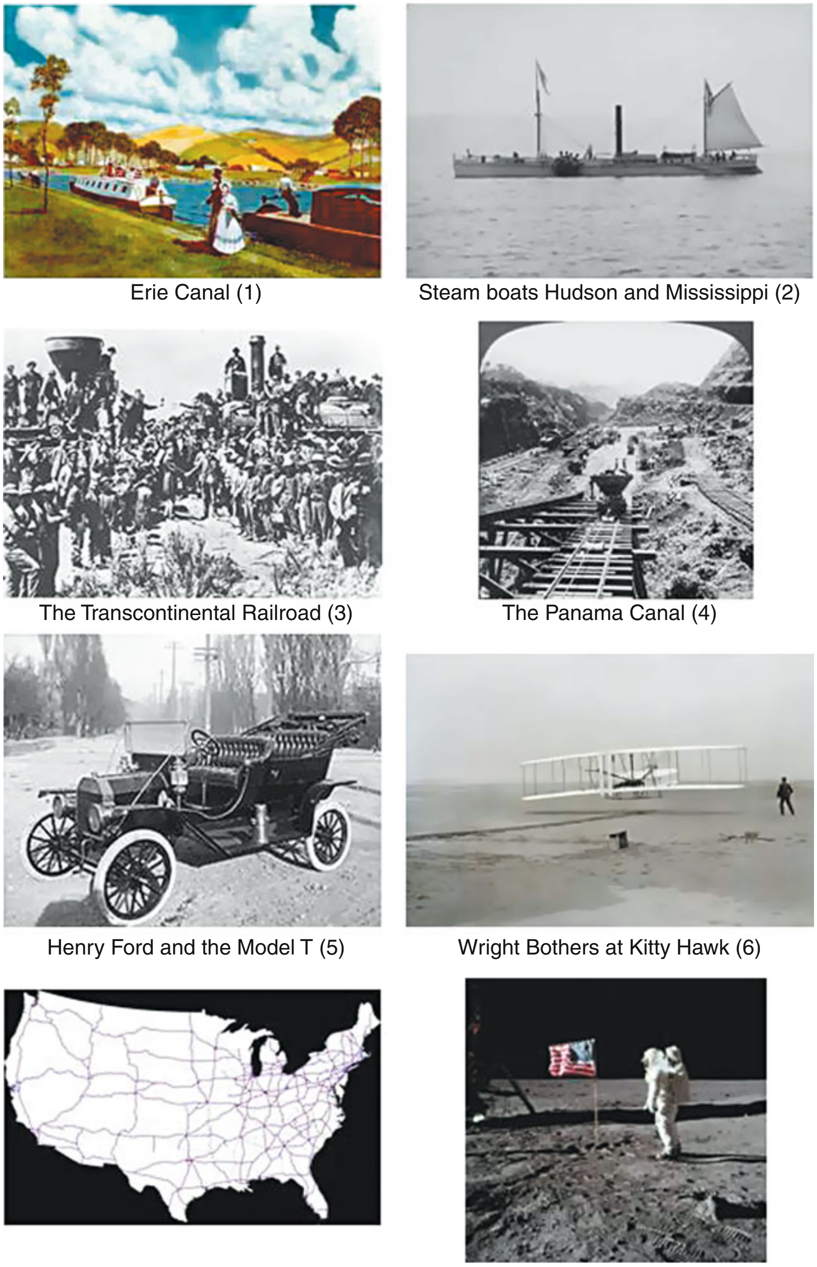
Over the last 200 years, the development of revolutionary new modes of transport have transformed human society, greatly raising the standard of living and enabling people to do things that were previously undreamed of.

Maglev, magnetic levitation, is such a revolutionary new mode of transport. It will dramatically change how humanity transports people and goods, launch payloads into space, store large amounts of electrical energy generated by wind and solar sources, and deliver billions of gallons of water over long distances to drought stricken regions. Moreover, it will be sustainable for the long term, and not require the use of destructive fossil fuels. In 1800, the only modes of travel were walking, horses, wagons, and sailing ships. In the USA, the per capita annual GDP was only $1,200 (in constant 2007 dollars) and the annual per capita travel distance was just 1,500 miles [1]. Today, the annual per capita GDP is $50,000 (in constant 2007 dollars) and the per capita annual travel distance is 15,000 miles [1].

What happened? A series of revolutionary new modes of travel that made it much faster, better, and cheaper to transport people and goods. During the Constitutional Convention, it took many days to travel from New England to Philadelphia, and to cross the ocean, many weeks. Even in the late 1800s, with steam boats and railroads, it took Phileas Fogg 80 days to go around the world.

The eight major transport revolutions initiated by the USA were (Fig. 1):

1. The Erie Canal (1814)
2. Steamboats on the Hudson and Mississippi (1807 and 1811)
3. The Transcontinental Railroad (1869)
4. The Panama Canal (1914)
5. Henry Ford and the Model T (1908)
6. The Wright Brothers and Kitty Hawk (1903)
7. The Interstate Highway System (1956)
8. The Moon Landing (July 20, 1969)



Erie Canal (1)

Steam boats Hudson and Mississippi (2)

The Transcontinental Railroad (3)

The Panama Canal (4)

Henry Ford and the Model T (5)

Wright Bothers at Kitty Hawk (6)

The Interstate Highway System (7)

The Moon landing (8)

MAGLEV Technology Development. Figure 1
The eight American transport revolutions

Maglev will be an equally transforming new mode of transport. On Earth, it will move people and goods without being powered by fossil fuels that emit massive amounts of greenhouse gases and pollutants, in vehicles that are very safe, efficient, convenient, quiet, and comfortable. Maglev will also store large amounts of electric energy at very high efficiency and low cost, transport very large amounts of water for long distances over difficult terrain, and launch payloads into space at very low cost, including space satellites that can beam large amounts of power to the Earth.

Why Fossil Fueled Transport Is Not Sustainable

Today's Transport Systems – autos, trucks, airplanes, most trains, and ships – are critically important to present society and its standard of living. All are critically dependent on oil. Without oil, humanity would be back in the 1700s, with only horses and wagons and sailing ships.

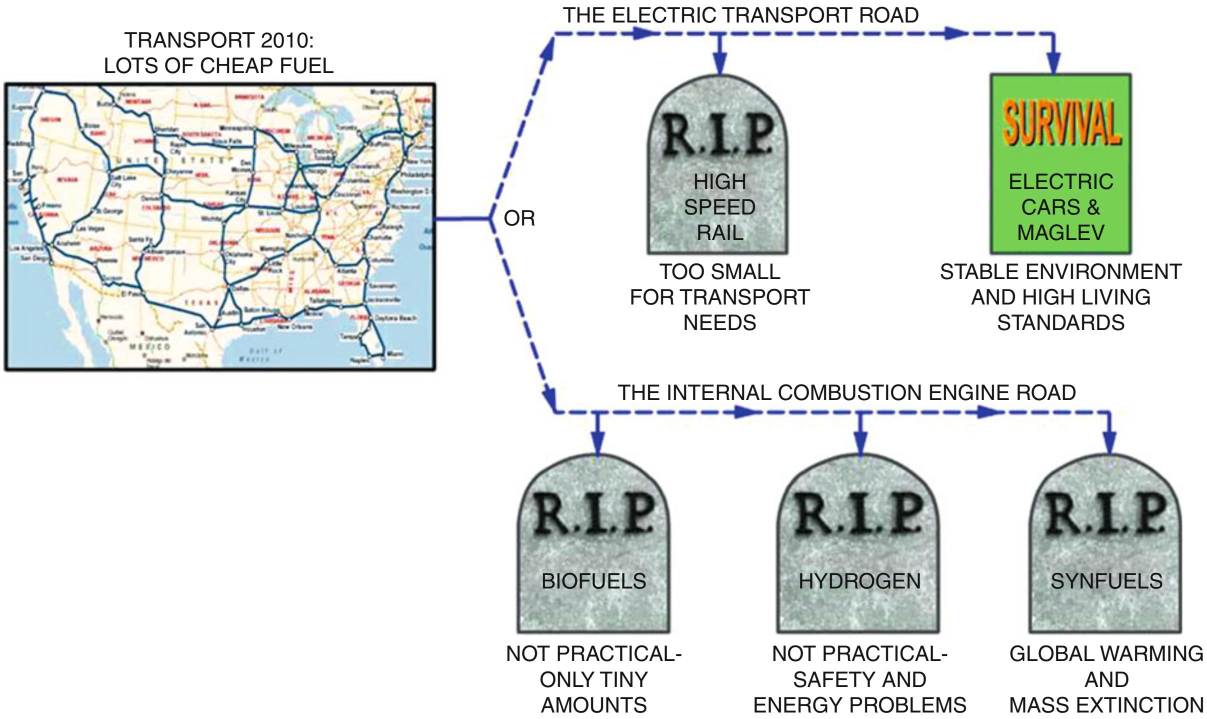
Since Edwin Drake started pumping oil in Western Pennsylvania in 1859, the world has consumed approximately one trillion barrels of oil, most of it for transport. How much is a trillion barrels? Visualize a lake of oil that is one square mile in cross section containing a trillion barrels of oil. The oil lake would be 40 miles deep.

Another trillion barrels of oil is still in the ground, but it will not sustain the world for very long. World oil production has plateaued at about 30 billion barrels per year and will soon start to steadily decline. More significant than the arguments over when the "peak oil" crisis will occur is the "oil competition" crisis. As countries like China and India rapidly industrialize, competition with the developed countries for ever scarcer oil will drive up the price of oil. Today, highway traffic in China is increasing 18% annually, and more cars are sold per year in China than in the USA. Today (2010) oil costs about $80 per barrel; in a few years, analysts predict $200 per barrel.

For most of the rest of the twenty-first century, the present world transport systems will not be sustainable, using conventional oil for fuel.

Can society sustain existing autos, trucks, airplanes, trains, and ships with some alternative fuel as the oil runs out? Biofuels? Hydrogen? Synthetic fuels derived from coal, oil shale, tar sands, methane hydrates from ocean beds? No. None of these will be able to maintain society's present transport systems powered by the internal combustion engine. There are a variety of reasons why these alternative fuels are not practical on the scale needed to support a world population of nine billion people in 2050 (Fig. 2).

Biofuels? Not practical for the amounts needed. The acreage required is prohibitively large. In a world


TRANSPORT 2010:
LOTS OF CHEAP FUEL
OR
THE ELECTRIC TRANSPORT ROAD
R.I.P
HIGH SPEED RAIL
TOO SMALL FOR TRANSPORT NEEDS
SURVIVAL
ELECTRIC CARS & MAGLEV
STABLE ENVIRONMENT AND HIGH LIVING STANDARDS
THE INTERNAL COMBUSTION ENGINE ROAD
R.I.P
BIOFUELS
NOT PRACTICAL- ONLY TINY AMOUNTS
R.I.P
HYDROGEN
NOT PRACTICAL- SAFETY AND ENERGY PROBLEMS
R.I.P
SYNFUELS
GLOBAL WARMING AND MASS EXTINCTION

MAGLEV Technology Development. Figure 2
The decision tree for future transport

where there is not enough arable land to adequately feed today's population of 6.6 billion people, which is projected to be 9 billion by 2050, biofuels can supply only a tiny fraction of future transport needs. Not only is the population growing, but the fertility of the world's arable land is declining, due to over use, erosion, increasing acidification due to excessive use of fertilizers, and increasing droughts. Humanity will be lucky to avoid massive famines in the years ahead, even if biofuels are not produced.

Hydrogen fuel would require 1,000 new nuclear reactors, each of 1,000 MW electrical capacity – ten times more than now operated in the USA – to make enough hydrogen fuel to equal the gasoline and diesel fuel currently consumed in USA's transport systems.

Hydrogen cars raise a serious security issue. Over one million cars are stolen in the USA per year. After stealing a hydrogen car, a terrorist need only attach a small bomb with an explosive-driven penetrator to the hydrogen tank, and then park it in a public parking garage, or shopping mall, or on a busy city street. The penetrator bomb could be then remotely detonated by cell phone or an attached timer, releasing the hydrogen gas to cause a massive explosion. Result? Many deaths with no way to prevent them. A terrorist would need no special skills. The penetrator bombs could probably be bought in the black market.

While more expensive than oil, synthetic gasoline and diesel fuel can be manufactured from coal, tar sands, oil shale, and methane at an acceptable cost. The reserves are enormous and could sustain oil-fueled transport for hundreds of years.

The problem with synfuels is not the sustainability of fuel supply, but sustainability of Earth's environment due to the release of carbon dioxide to the atmosphere from the manufacture of the synfuels, and the release of carbon dioxide from the tailpipes of the vehicles that burn the synfuels. Today, the US transport sector emits two billion tons per year of carbon dioxide into the atmosphere [2], roughly one-third of total US emissions. The carbon dioxide emissions from the manufacture of the synfuels needed to replace the oil now consumed by US transport would add another two billion tons per year to the atmosphere, for a total transport emission of four billion tons annually.

Would leaders call for a reduction of 80% in carbon dioxide emissions from today's levels, to be achieved by 2050? This corresponds to a total world emission of five billion tons annually in 2050. Four billion tons of just US transport emissions at today's level of transport would be four-fifth of the world's total.

With synfuels for transport, there is no way to reduce carbon dioxide emissions below today's level, even if all other present sources – coal-fired power plants, industry and home heating, etc. – were to go to zero. In fact, as the developing countries industrialize, their transport needs will greatly increase, making the world's total emissions from transport alone in 2050 much greater than today's total world emissions of 25 billion tons per year [3].

There is no practical way to capture the carbon dioxide emissions from autos, trucks, airplanes, trains, and ships. Inevitably, they will enter the atmosphere.

The climate effects of global warming are very bad – higher temperatures, melting glaciers and ice sheets, increased droughts and food shortages, rising sea levels and flooding, species shifts and extinction, etc. Even worse, at some point, high levels of carbon dioxide emissions risk environmental catastrophe that cannot be stopped, even if all emissions went to zero.

As global warming from the release of greenhouse gases continues, the ocean is becoming more acidic, threatening sea life. Natural carbon dioxide emissions from decaying organic matter in the permafrost regions, plus thermal decomposition of unstable methane hydrates in the sea bed, may cause an irreversible runaway warming of the Earth.

Humanity should not risk global disaster by continuing to emit massive amounts of carbon dioxide from its transport systems. A transition to electrically powered transport must soon start, with the electricity coming from nonpolluting energy sources (Fig. 2).

Maglev and Electric Transport: A Necessity for Long-Term Sustainability

There are three options: electric cars and trucks, electric rail, and Maglev. Electric cars are desirable and practical, but have limitations. The Chevrolet Volt, for example, which will soon be commercially available, has a maximum range of about 40 miles before it must be recharged or shifted to an auxiliary gasoline engine to extend its range [4]. Longer trips almost certainly will use the gasoline engine because nobody will want to stop every 40 miles or so and wait several hours for a recharge.

There are other limitations. In colder regions, batteries should not be charged or discharged unless they are warmed, as this may degrade them. Heating the car from the battery will reduce the car's range. In hot regions, running the air conditioner from the battery will also reduce range. The battery pack will be very heavy, hundreds of pounds, increasing the weight of the car.

In combination with Maglev, electric cars and trucks can provide practical transport, for both short and long trips. The electric cars can operate locally for trips of a few miles to the nearest Maglev station, there to drive onto a Maglev vehicle that will transport the drivers and their automobiles to any other station in the continental USA at speeds of 300 mph. At the station nearest the destination, the car would be driven off the Maglev vehicle, a few miles by highway to the final destination.

During the trip on the Maglev vehicle, the electric cars would be fully recharged, using on-board power, and ready to travel once driven off. The cost to drivers traveling on a high speed Maglev trip would be less than if driving by highway with a gasoline-powered engine. Maglev would provide the same kind of service for long distance trucks. The average haul distance for intercity trucks, which currently transport 1½ trillion ton miles of freight per year at an annual cost of over $300 billion, is 500 miles. There is no way that electric trucks can make that haul distance if traveling by highway.

With a National Maglev Network, almost all travel could be electrically powered, with long distance travel for autos, trucks, and passengers by Maglev, while electrically powered autos and trucks handled short local trips. There would be little domestic air travel, with most flights being across the oceans and to remote areas not served by Maglev.

Electrically powered conventional rail will still transport bulk freight – grain, cement, etc. – but, the role of high speed rail (HSR) will be limited. This is because HSR only carries passengers, is inherently expensive, and must be subsidized by the government, even in Europe and Japan. Even in countries like France which is much smaller than the USA, with considerably high population density, for example, HSR plays only a small role in transport [1]. Per capita, on average the French travel only 400 miles per year on the TGV, while driving 7,600 miles per year and only taking a 1¼ trip per year on the TGV.

The Physics and Engineering of Maglev

The Physics of Maglev

It has been well known for well over a century, since the beginning of the electrical age, that electromagnetic forces can be very strong, even stronger than gravity. This principle makes magnetic levitation possible. Moving charged particles, which usually means electrons, create magnetic fields. Matter contains atomic electrons perpetually whirling around in orbits. Most materials are largely nonmagnetic because the moving electrons cancel each others' magnetic fields, even when an external magnetic field is applied. A few materials are ferromagnetic. This material responds to external magnetic fields by aligning the electron orbits to produce very strong induced magnetic fields within the ferromagnet that extend beyond it into the surrounding air. When the external magnetic field is removed, the ferromagnetic object essentially returns to the unmagnetized state. It is fortunate that iron, a common and cheap structural material, is strongly ferromagnetic. This enabled the creation of the generators and transformers that made AC electric systems very practical.

It was quickly realized, however, that stable suspension in air is not possible with ferromagnetic materials using static, that is, non-time-varying, external excitation. This is known as Earnshaw's Theorem. The unstable force causes magnetized iron to clamp itself to another piece of iron with a force which gets stronger as it gets closer. This is clearly not suitable for levitation.

Permanent magnets are familiar to people and can be used for levitation. These magnets are made of unusual materials which resist easy magnetization or demagnetization. When exposed to very strong external magnetic fields during manufacture, a large fraction of the electron orbits line up parallel and retain strong magnetization alignment after the charging magnetic field is removed.

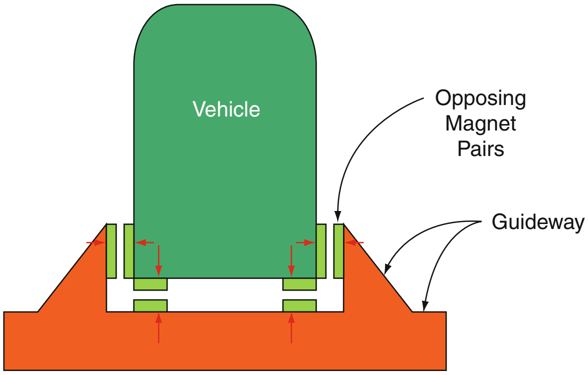
Two such permanent magnets will attract each other when the magnets' electron orbits are parallel. However, if one magnet is turned over, the magnets will strongly repel each together. As the magnets get closer, the repulsive force becomes stronger. At some distance, the repelling force is sufficiently strong to equal that of gravity.

This then leads to suspension at a height where the repulsion force equals that of gravity. However, without additional constraint, this type of suspension is not practical for levitation. The suspended magnet is unstable at right angles to the direction of gravity. A familiar example is two donut-shaped permanent magnets centered on a stick. With the magnetic polarity opposing, the lower magnet will suspend the upper magnet. However, if the lower magnet is moved up on the stick to the point where the upper magnet leaves the top of the stick, the latter will be violently thrown off to the side.

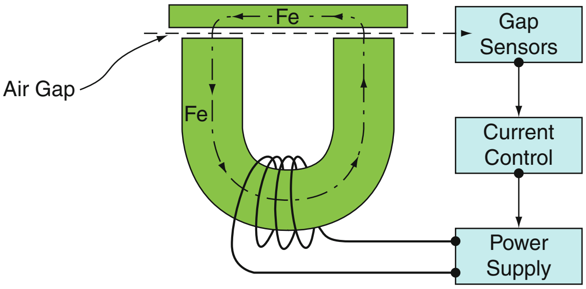
Figure 3 shows how stable levitation can be produced. This is a familiar Maglev model construction for school science projects. The side mounted magnets and the vertically mounted magnets in combination create forces which are highly stable in both directions. For a real Maglev transport system, low cost familiar ferrite magnets can only levitate a very small weight. Moreover, to build many miles of permanent magnet guideway would require an enormous volume of very expensive material to be purchased, installed, and maintained. There is a class of Rare Earth permanent magnets which produce forces at least ten times greater than ferrite magnets. However, this material which weighs as much as iron costs hundreds of times more per pound than iron. In summary, for a robust Maglev system with substantial suspension clearances that can levitate heavy weights, using a permanent magnet approach is not practical.

Let us return to the ferromagnetic circuit which is statically inherently unstable. Figure 4 illustrates that if the applied magnetizing current in a coil is electrically varied over time, the unstable force can be controlled so that a practical suspension is possible. Sensors continually monitor the air gap and signal the power supply to adjust the circuit current in a suitable time-dependent manner so that the air gap is maintained at a constant value. This is the system used by Transrapid [5]. As was the case for permanent magnets shown in Fig. 3, both horizontal and vertical magnets are required for stability in both directions. On each side of the guideway, magnets attached to the vehicle are attracted upward to the bottom of the iron rails on the guideway, providing levitation. Magnets of each side are attracted sideways to the edge of the guideway, as illustrated in Fig. 5, providing lateral guidance. The operating gap between the vehicle magnets and the iron rails on the guideway is typically very small with the electromagnet iron rail system, on the order of 1 cm (3/8 in.).

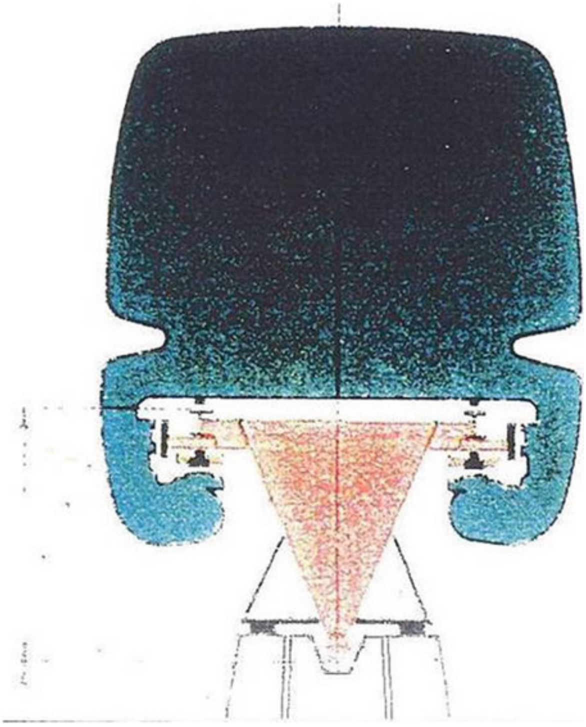
Another type of magnetic suspension results from applying alternating current (AC) to a current-carrying coil located above a sheet of electrically conductive normal metal, such as copper or aluminum. If the frequency is sufficiently high, currents are induced in the metal sheet that keeps the AC magnetic field from



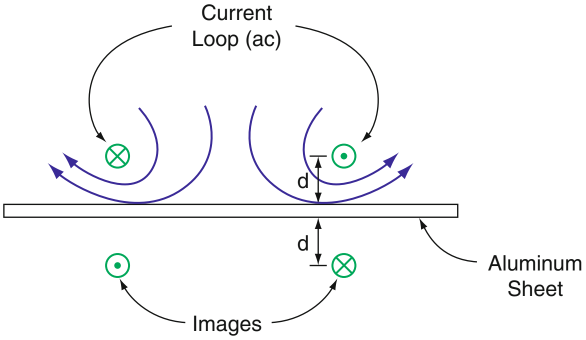
MAGLEV Technology Development. Figure 3 Magnetic suspension based on the repulsive force between permanent magnets on a vehicle



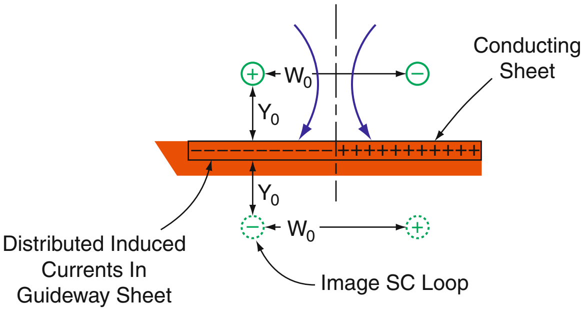
MAGLEV Technology Development. Figure 4 Magnetic suspension based on servo control of the attractive force between an electromagnet and a ferromagnetic sheet



MAGLEV Technology Development. Figure 5
Cross section view of Transrapid Maglev vehicle showing electromagnetic suspension



MAGLEV Technology Development. Figure 6
Magnetic suspension based on the repulsive force between a coil carrying alternating current (AC) and a conducting sheet



MAGLEV Technology Development. Figure 7
Superconducting (SC) inductive suspension conducting sheet guideway

penetrating the sheet. These induced currents are opposite in direction to the currents in the AC coil above the sheet. The magnetic interaction between the opposing currents produces a repulsive force similar to that of opposing permanent magnets which can levitate the AC coil (Fig. 6). As early as 1912, Batchelet [6] obtained a patent on such a system for vehicle transport. While not practical for transport because of its very inefficient use of energy, it does work from the physics viewpoint.

The concept of induction brings us to electrodynamic Maglev. Note that induction occurs because of time-varying applied magnetic fields. The usual example is an applied AC current, for example, as seen in transformers, which make our electric utility power distribution practical. In the case of Maglev, static (DC) magnets on a vehicle moving along the guideway can also apply a time-varying, that is, AC field to normal conductors that they pass on the guideway.

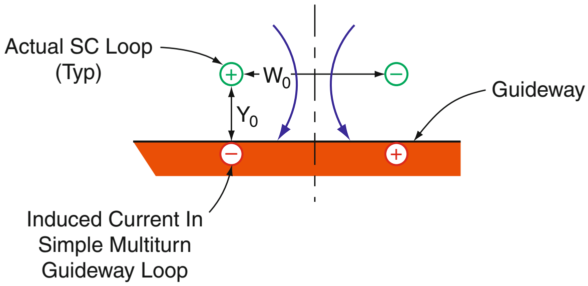
Powell and Danby shared a house from 1960 to 1962. This was about the time superconducting wires capable of producing strong superconducting magnets were first becoming available. Dr. Powell had studied the use of superconductors positioned on a guideway and the vehicle for Maglev. Then Dr. Danby provided key electrodynamic insights into using induction from superconducting magnets on a vehicle to generate currents in normal conductor loops on the guideway. Together Danby and Powell proceeded to develop and patent superconducting electrodynamic Maglev for which the Benjamin Franklin Medal for Engineering in April 2000 was awarded [7].

These innovations are described on the following figures. Figure 7 shows a superconducting coil moving above a conductive sheet. This configuration, while possible, is less desirable, but is readily visualized.

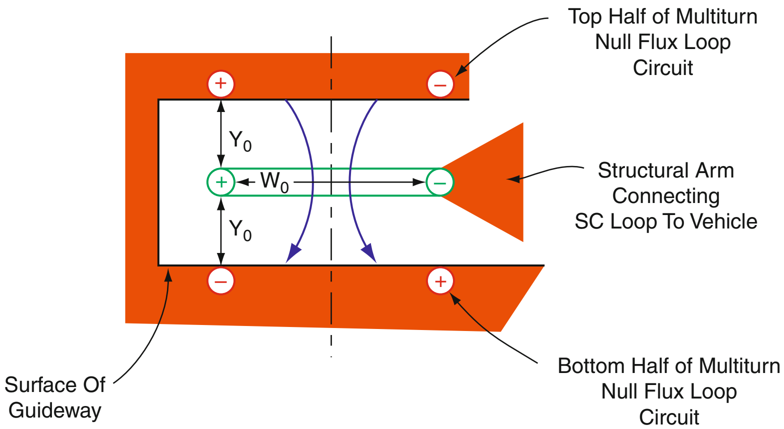
Figure 8 shows the conductive sheet replaced by a bundle of wire forming a shorted coil. Figures 9–13 illustrate a variety of the many configurations for use of what the authors' term "null flux" loop geometries. These null flux configurations, as discovered by the authors, have proved to be a very desirable approach for superconducting electrodynamic Maglev. The use of strong superconducting magnets enables the vehicle to levitate very heavy loads while at the same time keeping electrical I^2R losses in the normal guideway conductors at a minimum level [7, 8].

In Fig. 9 the superconducting (SC) magnet loop, which is attached to the vehicle, is shown halfway between the guideway loops. The induction in the top and bottom guideway loops is equal and opposite, since these loops are added in series opposing. As a result, no current flows, no force is exerted, and no power is consumed. As the vehicle magnet drops down because of the vehicle weight, a current and force proportional to displacement occurs. The vehicle is suspended with a restoring force as in a spring. In Fig. 10, two loops on the guideway are mounted side by side and connected in series opposition. If the SC loop on the vehicle is centered, there is no net induction or current or force. This gives a strong linear restoring force sideways as the vehicle moves sideways off center. Figure 11 is identical to Fig. 10, but oriented at 90°. Now the force is vertical instead of horizontal. Figure 12 shows another example. When centered there is no induction or sideways force due to the guideway loop. Figure 13 shows canceling of the force between the two sides when the vehicle is centered. In all the above cases, strong restoring forces are developed proportional to the displacement from the ideal position of the vehicle, that is, centered. The figure of 8 configuration is used in the very successful Japanese Maglev [9].

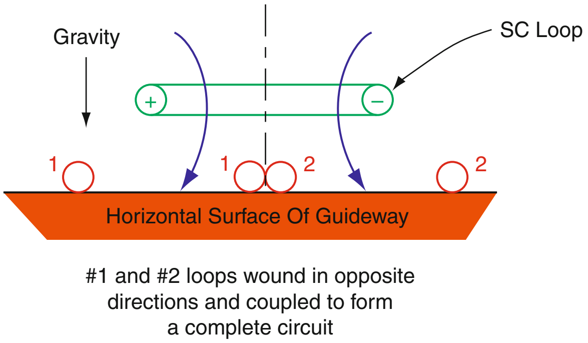
Why is the null flux principle so powerful? For the simple loop of Fig. 8, the geometry and the total ampere turns in the superconducting magnet completely determine the induced current in the shorted guideway loop. With the null flux configuration, the number of ampere turns in the vehicle SC magnets can be much greater than the guideway ampere turns and still have a very strong suspension.



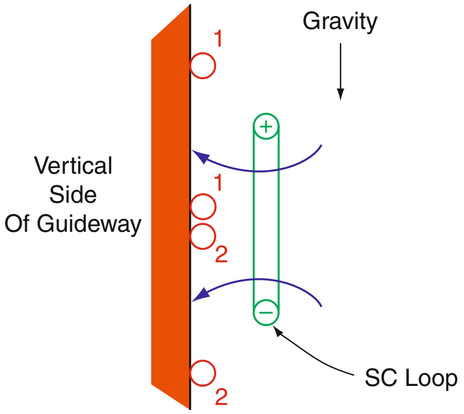
MAGLEV Technology Development. Figure 8
Superconducting (SC) inductive suspension simple loop guideway



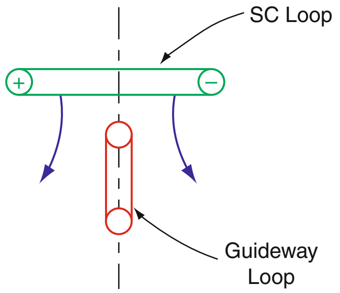
MAGLEV Technology Development. Figure 9
Series opposed loops


Gravity
SC Loop
+
−
1
1
2
2
Horizontal Surface Of Guideway
#1 and #2 loops wound in opposite directions and coupled to form a complete circuit

MAGLEV Technology Development. Figure 10
Horizontal figure of 8



MAGLEV Technology Development. Figure 11
Vertical figure of 8



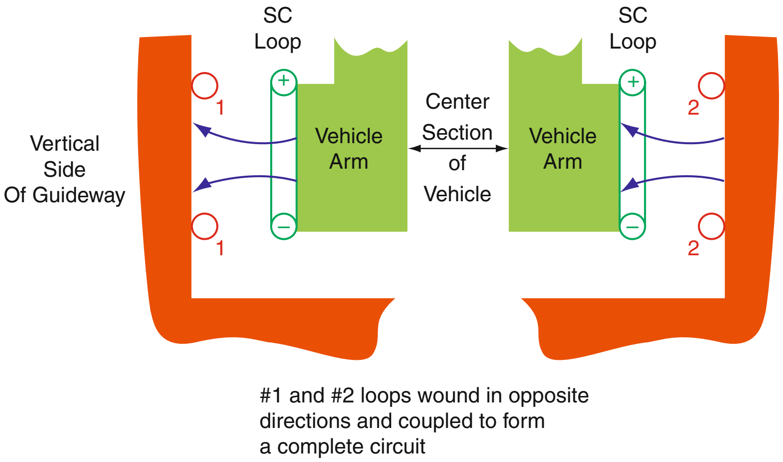
MAGLEV Technology Development. Figure 12
Central orthogonal loop (horizontal or vertical orientation)

This reduces material requirements for the guideway coils and results in low I^2R losses in the guideway. Stronger SC magnets on the vehicles are well within the state of the art. These stronger magnets in turn result in less power needed for the linear synchronous motor discussed later. Moreover, because of the very efficient null flux levitation configuration, most of the propulsion power required for Maglev vehicles at high speeds is due to air drag, not electrical loss.

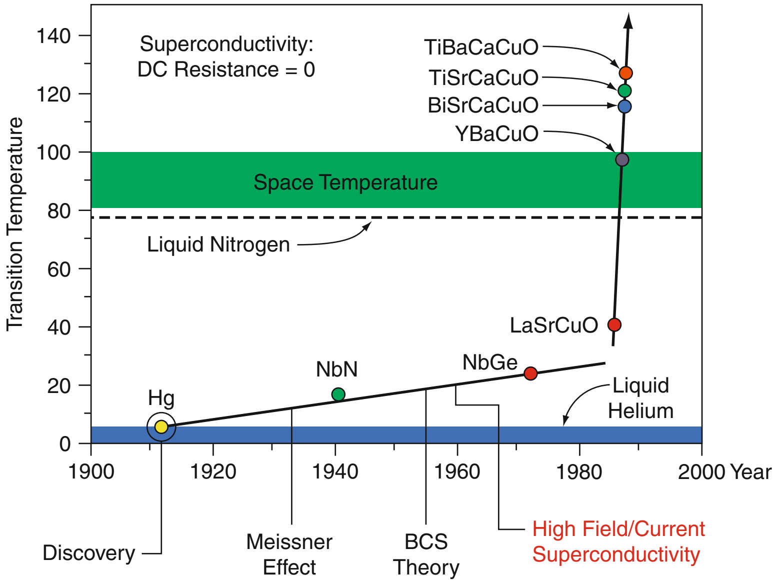
Throughout the history of technological evolution, ideas that were known in principal had to await new technical advances to make them truly practical. That is the case with Maglev. Early visionaries of Maglev over the last 100 years include Bachelet [6], Goddard [10], and Kemper [11]. Lightweight and strong superconducting magnets provided a great step forward for electrodynamic Maglev. The materials that made this possible appeared about 1960. This in combination with the authors' null flux loop guideway made Maglev eminently practical. Similarly, the development of modern power supplies, controls, and sensors permits fast-acting electronic control of the iron magnets of the Transrapid system which has been demonstrated in Germany and commercially implemented in China.

From the Maglev point of view, a superconducting winding behaves like a very light and strong permanent magnet which is located in a very low temperature container, or cryostat. Superconductivity is now in routine use, especially in the MRI scanners found in hospitals and doctors facilities. Figure 14 shows the history of superconducting development. At the very low temperatures of liquid helium, 4.2 K, Kammerlinge Ohnes [12] discovered that various pure metals had no resistance to current flow. In turn, the flowing electron current generates magnetic fields without the I^2R losses experienced in normal conductors. While the discovery was of great interest to physicists, it was almost 50 years before superconductors were developed which could carry large currents in the presence of high magnetic fields as necessary for magnets. The most common superconducting wire for magnet builders for the last several decades has been Niobium-Titanium (NbTi) followed by Niobium 3-Tin (Nb_3-Sn).

Figure 15 shows the behavior of NbTi alloy. The lower the operating temperature, the greater the superconducting current and magnetic field that can be sustained, that is, the more powerful the magnet capability. Figure 16 shows the behavior of these two most common superconductors at the temperature of boiling helium. Note the much greater current densities

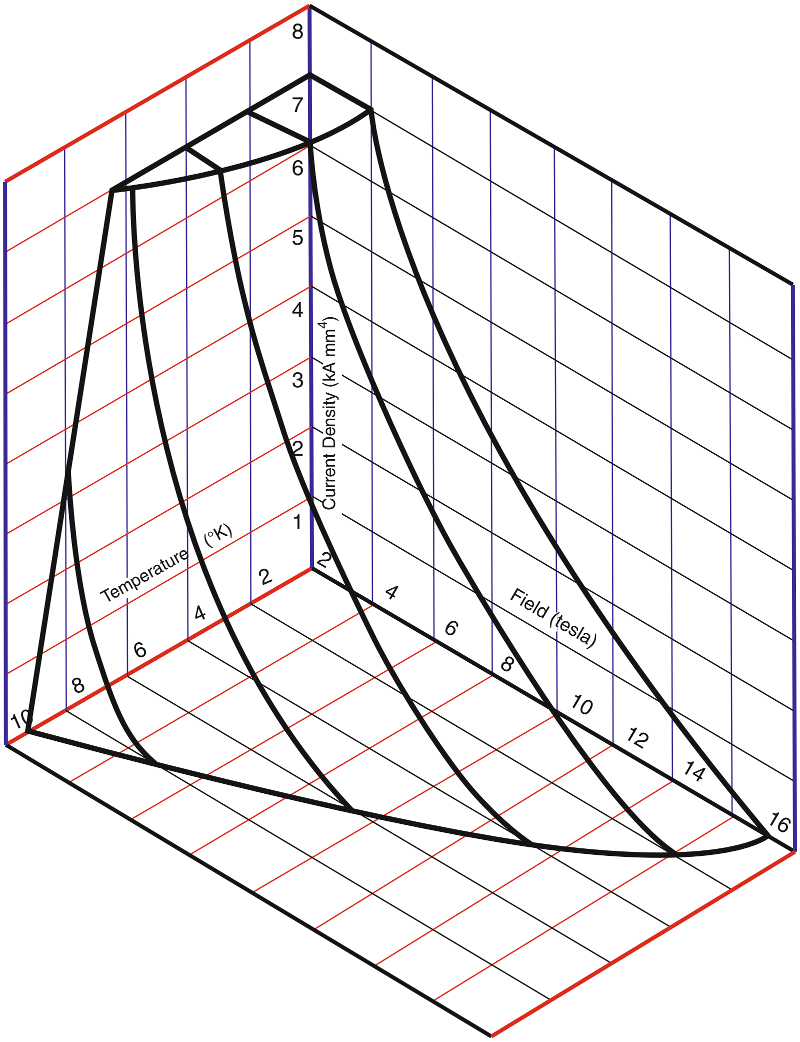


MAGLEV Technology Development. Figure 13
Cross-coupled lateral loops



MAGLEV Technology Development. Figure 14
The time evolution of the superconducting transition temperature

and resulting magnetic fields that can be sustained compared to conventional iron electromagnets. Figure 17 shows NbTi wire drawn down to a circular or an approximately rectangular form. The small dots are the NbTi superconductor. The surrounding material is usually copper. This cross section starts out as a short cylinder about 10 in. in diameter. It is then drawn down to a small wire by pulling through orifices of diminishing sizes. The superconducting wire and cryostat technology required for Maglev has existed in the USA for several decades. For example, Fig. 18a shows a large superconducting magnet system which was operated for over a decade as a central part of a high energy physics facility. It was designed and

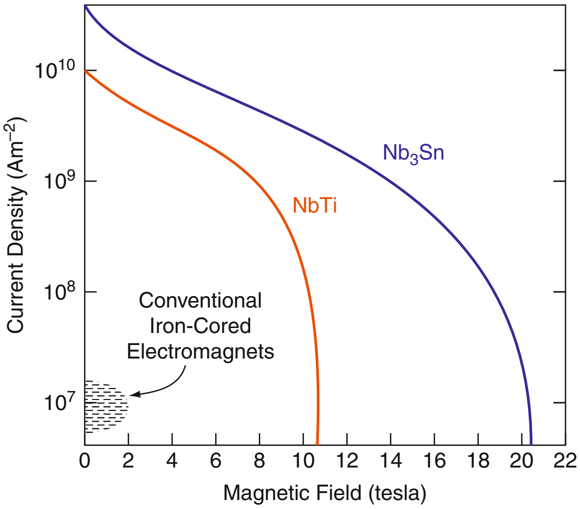


MAGLEV Technology Development. Figure 15
The critical surface for niobium-titanium superconductor

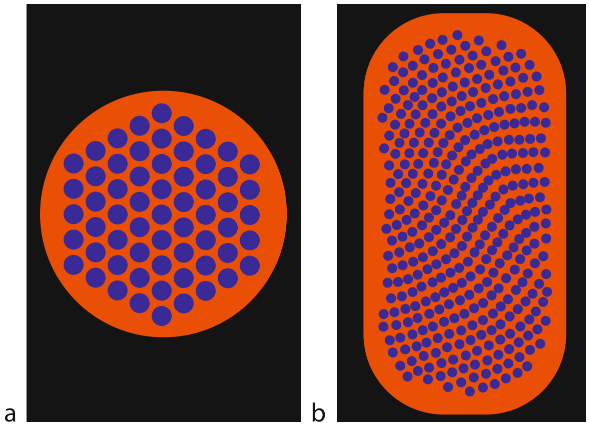
built by Gordon Danby and his group and commissioned in 1973. Figure 18b shows a large superconducting magnet system used as a high energy particle storage ring for a fundamental physics study of the anomalous magnetic moment of the muon, a fundamental point-like particle.

The largest and longest particle accelerator that uses superconducting magnets is the Large Hadron Collider (LHC) now operating at the CERN Laboratory in Switzerland [13]. The LHC has two rings of many thousands of superconducting magnets that accelerate and guide high energy particles for experiments. The total circumferential length of the magnets is 54 km (31 miles). The same length of superconducting magnets on Maglev vehicles could transport half of the 4 trillion passenger miles per year traveled by car, airplane, and rail in the USA, plus half of the 1.5 trillion ton miles per year of freight transported by intercity highway trucks.

The LHC superconducting magnets are much more technically complex and challenging than superconducting Maglev magnets. To achieve maximum field capability, LHC superconducting magnets operate with subcooled liquid helium at 1.7 K, rather than the normal boiling point of 4.2 K, where Maglev magnets operate. The refrigeration requirements at



MAGLEV Technology Development. Figure 16
Critical current densities of niobium-titanium and niobium-tin at a constant temperature of 4.2 K



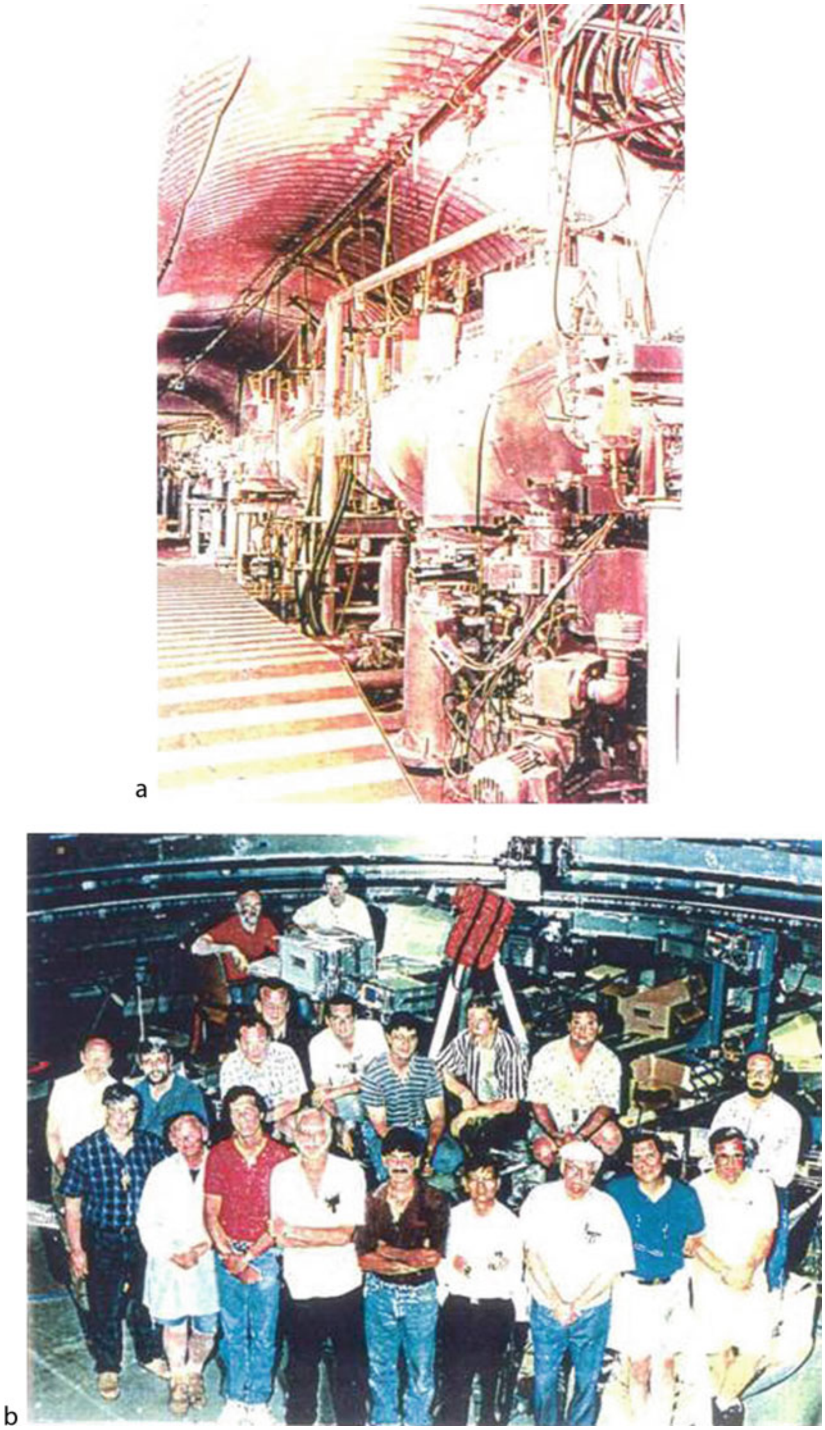
MAGLEV Technology Development. Figure 17
Cross sections of two typical filamentary composites of NbTi in a copper matrix

1.7 K are much more difficult than at 4.2 K. Moreover, the LHC superconductors must operate at higher magnetic field strengths and with much greater spatial precision than Maglev magnets again making them much more challenging. Finally, the many thousands of LHC magnets must all work perfectly for the LHC facility to operate. In Maglev, the vehicles carry multiple independently operating magnets with very high system redundancy and robustness. If one magnet were to fail, a very unlikely event, it would not affect the operation of the other magnets on the vehicle. Several of the multiple independent magnets on the vehicle could fail without compromising safety. If a magnet were to fail, the vehicle would be taken off-line at the next station for repairs. The probability of multiple failures during this time would be many orders of magnitude smaller than the probability of multiple engine failures on commercial jet aircraft, which in turn are extremely small.

Returning to the cost of the magnets for superconducting Maglev, a good measure is the cost of the magnets for the LHC. The total capital cost of the much more complex and difficult LHC magnets was several billion dollars. The equivalent length of magnets on Maglev vehicles would carry half of the annual US passenger and intercity highway freight traffic, which presently costs on the order of $1 trillion per year [14]. The amortized cost of the Maglev magnets is trivial in comparison. Also, the annual cost of the 40,000 deaths and 500,000 serious injuries per year on US highways is estimated at $500 billion annually [1]. The cost of superconducting Maglev magnets is again trivial by comparison, plus the vast amount of human suffering that would be avoided by traveling on much safer Maglev systems.

The newer high temperature superconductors (HTS) shown in Fig. 14 are extremely promising. HTS conductors operate at much higher temperatures than the 4.2°K for NbTi, typically in the range of 30–40 K, with the capability of operating at 77 K, liquid nitrogen temperatures, depending on the desired current capacity. Presently, the most promising HTS conductor is yttrium barium copper oxide (YBCO). Deposited as a very thin film, 1–2 μ in thickness, on a thin tape substrate, YBCO conductors can achieve engineering current densities of 50,000 A/cm^2 or more, depending on operating temperature. Engineering current density is the overall current density of the conductor tape, including the superconductor film, which has a current density of millions of amperes per square centimeter, plus the substrate tape that it is deposited on.

HTS conductors are attractive for Maglev, because their refrigeration systems are simpler and have lower energy requirements. The present Japanese Maglev



MAGLEV Technology Development. Figure 18
(**a**) Large superconducting magnet systems for a high energy physics facility. (**b**) Large superconducting magnet system used as a high energy particle storage ring for study of the magnetic moment of muon particles

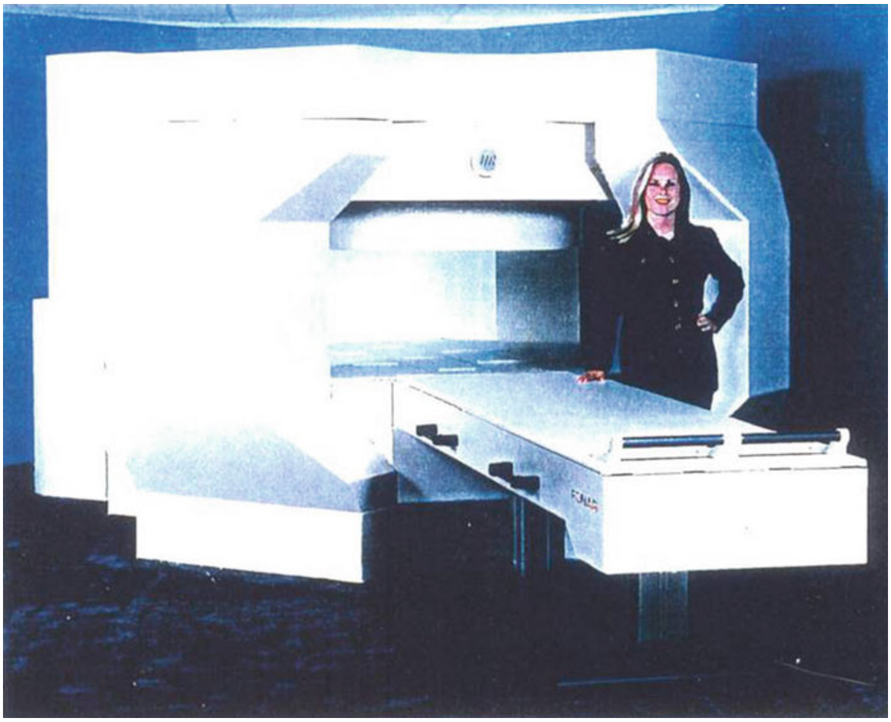
system uses NbTi superconductor at liquid helium temperatures without problems, and the refrigeration requirements are modest and very acceptable. As HTS conductors come into wider use and larger production, superconducting Maglev will probably shift to their use.

A further question could be raised by people unfamiliar with superconducting technology. Is it too complex for everyday transportation systems? One area where superconductivity is intensively used is MRI medical scanners. The technology of superconducting coils and cryogenics has evolved considerably in scanners over the last 20 years, and thousands of superconducting MRI magnets are now in use all over the world. Figure 19 shows an open MRI, for which Gordon Danby was the leader of the magnet design effort. The important thing for Maglev is that the commercially made superconducting coils in the MRI scanner as yet require elaborate refrigeration systems. Two small cryocoolers approximately a foot in diameter and an arm's length cool the coils. Once a year the cryocoolers are taken off and replaced with another set. The first set is sent back to the manufacturer for reconditioning. No cryogenic fluid has to be added during this year of continuous trouble-free operation. This example shows that superconductivity can be very reliable, just like the refrigerators in people's kitchens. It is definitely ready for efficient, reliable transportation with extremely low maintenance.

Engineering Maglev

If Maglev is to become a major mode of transport in the twenty-first century, as important as road, rail, air, and water have been in the twentieth century, it must be safe, reliable, cost effective, environmentally benign, operate as a nationwide network, not a small number of isolated routes, be readily accessible, and integrate easily with other transport modes. To fulfill Maglev's promise, the proposed Maglev systems must be engineered with respect to the following six development goals.

The first development goal is to have a substantial physical clearance between the Maglev vehicle and the



MAGLEV Technology Development. Figure 19
Magnetic Resonage Imaging (MRI) machine for medical diagnostics, using superconducting magnets

guideway. Large clearances are desirable for safe operation, and allow Maglev vehicles to operate in all types of weather, including strong winds, snow, and ice, that are not possible with conventional transport. Superconducting Maglev enables clearances of 10 cm (4 in.) or more between a vehicle and the guideway. Electromagnetic Maglev is constrained to much smaller clearances, on the order of 1 cm (3/8 in.). Also, large clearances reduce guideway cost because construction tolerances are less demanding. Building guideways to a tolerance of ±10% of the vehicle/guideway clearance is much cheaper when the clearance is 4 in., compared to guideways where the clearance is a fraction of an inch.

The second development goal is to have a very stable and safe magnetic suspension. Electromagnetic Maglev systems, such as Transrapid, achieve stability by very fast, for example, thousandths of a second, servo control of the current in the electromagnets on the vehicle. If the gap between the vehicle electromagnets and the iron rails becomes smaller than desired, the magnet current is reduced. If the gap becomes too large, the magnet current is increased. In electrodynamic superconducting Maglev, the vehicle superconducting magnets and the normal conductor guideway loops are arranged so that the moving vehicle will automatically be magnetically suspended at an equilibrium position above the guideway. Any external force on the vehicle – for example, a cross wind or the centripetal force on a curve – that acts to displace the vehicle from its equilibrium position is immediately countered by an inherent, passively generated opposing magnetic force. This opposing magnetic force always acts to push the vehicle back toward its equilibrium position. The magnitude of the displacement from equilibrium is proportional to the strength of the external force. Electrodynamic superconducting Maglev can be designed so that no conceivable external force, even if it were twice the weight of the vehicle, would produce a displacement great enough to make the vehicle contact the guideway. In effect, superconducting Maglev creates a five-dimensional "magnetic well" in which the vehicle is trapped so strongly that any conceivable external force cannot make it climb out of the "well." The five dimensions correspond to vertical ("heave"), lateral ("sway"), pitch, yaw, and roll movements of the vehicle. The vehicle is not constrained in the sixth, or longitudinal, dimension, so that it can move freely along the guideway, with only a small magnetic drag force acting on it.

Electromagnetic Maglev systems, that is, Transrapid, maintain their levitation at all speeds, including zero speed at stations and if the propulsion power to the guideway is cut off, subject to the electrical capacity of the batteries on-board the vehicle.

Superconducting Maglev vehicles can automatically and inherently remain stably suspended as long as they move faster than a transition speed of approximately 20 mph. Below this speed, resistive IR losses in the guideway loops reduce the induced currents in the guideway loops to the point where the magnetic force is insufficient to levitate the vehicle.

Superconducting Maglev vehicles can remain levitated at zero speed and speeds below 20 mph on the short sections, for example, 50 m in length, of guideway at stations where the vehicle is decelerating into a stop at the station or accelerating out of the station. To remain levitated in those special sections, one can power the aluminum guideway loops with DC current or use high temperature superconducting (HTS) guideway loops in place of aluminum loops. The HTS guideway loops will only require a small fraction of the HTS conductor used on the vehicle, and their refrigeration power will be small. Maglev vehicles are magnetically propelled by a set of alternating current propulsion windings in the guideway. In the unlikely event that the electrical power grid that energizes the guideway propulsion windings was to fail, vehicles would simply coast to a safe stop at a special location on the guideway or at the next station where the vehicle remains levitated.

The third development goal is to have the vehicles operate with very low magnetic drag. Low magnetic drag reduces both the energy consumed by the magnetic propulsion system, as well as its capital cost. Because the energy consumed by the superconducting magnets on the vehicle is negligible – the electrical resistance of superconducting windings is zero – low magnetic drag can be achieved by having the magnitude of the currents induced in the guideway loops be much smaller than the magnitude of the currents in the superconducting magnets on the vehicle. The magnetic levitation force on the vehicle is proportional to the product of vehicle magnet current and guideway loop current. Maximizing vehicle current relative to guideway loop current minimizes magnetic drag since it minimizes the I^2R losses in the guideway loops, the only contributor to magnetic drag. By using a unique configuration of guideway loops and vehicle magnets, termed the null flux suspension, to be described later, the guideway currents can be made very small, resulting in magnetic drag forces that are negligible compared to air drag.

Following publication of Powell and Danby's pioneering 1966 paper [15] on superconducting electrodynamic Maglev, Kolm and Thornton proposed a different Maglev system, called the Magneplane, that also used superconducting magnets on a vehicle to provide levitation and propulsion. Instead of using null flux aluminum guideway loops, the Magneplane [16] used a curved solid aluminum sheet as the guideway underneath the Maglev vehicle. The magnetic fields from the superconducting magnets on the vehicle induced currents in the aluminum sheet guideway. The magnetic interaction between the induced currents in the aluminum sheet guideway and those in the superconducting magnets on the vehicle generated the levitation force.

However, in this configuration, the magnitude of the induced currents in the sheet guideway equaled the magnitude of the currents in the vehicle magnets, resulting in large I^2R losses in the aluminum sheets and strong magnetic drag in the vehicle. Passenger aircraft typically have a lift to drag ratio on the order of 20/1, resulting in high energy consumption per passenger mile. The magnetic lift/drag ratio for the Magneplane is comparable to that for airplanes, again resulting in high energy consumption.

In comparison, by using the null flux guideway configuration [7, 15], the induced currents in the guideway loops are much smaller than those in the vehicle magnets, resulting in much less magnetic drag due to I^2R losses in the in the guideway conductors, and a much larger magnetic lift/drag ratio. Also, the amount of aluminum used in the null flux configuration is much less than in the Magneplane configuration.

The fourth development goal is low capital and operating costs. Most – approximately 80% – of the projected total life cycle cost of a Maglev system is associated with the construction cost of the guideway. A low cost for the guideway appears feasible through the use of prefabricated beams and piers which are mass produced in factories and shipped to the route site where they can be easily assembled into the finished guideway. Not only will the prefabricated, factory-produced beams and piers be substantially lower in cost than if they were fabricated at the construction site, but their quality control and tolerances will be better. The prefabricated guideway loops can be attached to the sides of the beams before they are shipped to the route site, which further reduces cost. The finished prefabricated beams and piers can be shipped by truck, or transported along completed portions of the guideway to where the new sections are being assembled.

The fifth developmental goal is to achieve high revenues from the Maglev system, so as to avoid the need for government subsidies for construction and operation of Maglev routes. Today, as government budgets are tightening and deficits and debt are growing, there is great resistance to having government finance new transportation projects. In the USA, intercity and commuter rail systems all require large subsidies for operation, with fare revenues only accounting for 30–40% of total operating cost. Similarly, the high speed rail (HSR) passenger transport systems in other countries all require major governmental subsidization. For Maglev to be widely implemented, it should require little or no subsidization, and as a goal, be attractive to private investment, without the need for any government funding.

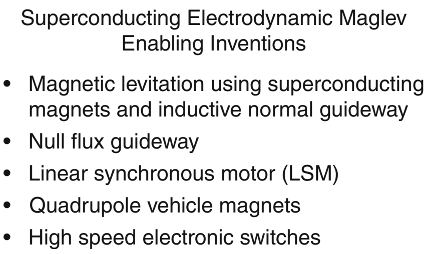
The sixth developmental goal is high speed, nonmechanical switching to off-line stations for unloading/loading operations. Requiring high speed vehicles to operate with online stations greatly limits their convenience and number of stations. Decelerating into and accelerating out of stations greatly reduces the average travel speed unless the stations are far apart, for example, spaced every 50–100 miles or so. Long distances between online stations reduce accessibility and ridership, while short distances negate the benefits of high speed systems. The ability to switch at high speeds to closely spaced off-line stations, when scheduled, and to bypass them at high speeds when not scheduled, enables a high speed transport system to maintain high average speeds while at the same time having many closely spaced stations with easy accessibility.

The nonmechanical contact nature enables a unique new capability not possible with rail systems. Suitably designed, Maglev vehicles can electronically switch at high speeds without the need for mechanically moving long sections of the guideway. Not only does electronic switching avoid the cumbersome mechanical movement, it enables switching at much higher speeds than if the vehicle had to slow down to navigate the mechanical switch section.

Figure 20 lists the superconducting electrodynamic Maglev inventions pioneered by the authors that address the six development goals outlined above.

The first invention is the use of superconducting magnets on the vehicle to induce currents in normal metal – for example, aluminum – guideway loops to levitate and automatically stabilize high speed vehicles [7, 15]. This combination efficiently uses the very strong magnetic fields generated by a small number of superconducting magnets on the vehicle that interact with simple, low cost, low drag normal aluminum guideway loops to levitate the moving vehicles. The combination is ideal, since using normal electrical conductors on the vehicle would require an impractical amount of on-board power for the large clearances involved, and superconducting loops on the guideway would be far too expensive.

The second invention, illustrated in Fig. 21, is the null flux guideway. The principle of the null flux guideway is simple. The guideway loops are arranged so that they have symmetry positions relative to the vehicle magnets. When a vehicle magnet is at the symmetry position, the *net* magnetic flux through the given guideway loop or loop circuit is *zero*. (The null flux condition can be designed to apply to either a simple loop or a circuit of two or more connected loops.) When the net flux through the null flux loop or circuit is zero, the resultant induced current in the loop or circuit is also zero. If the vehicle magnet moves away from its symmetry position, the *net* flux through the loop or circuit becomes *nonzero*, and a current flows that acts to push the vehicle magnet back toward its symmetry position. A wide variety of loops and loop circuits can be used in the null flux guideway. A very useful configuration is the figure of 8 null flux loop circuit, as illustrated in the upper right corner of Fig. 21. The left-hand loop of the figure of 8 circuit is wound in a clockwise direction, while the right hand loop is wound counterclockwise. When the dipole loop vehicle magnet is centered on the figure of 8 loop circuit, the net flux through the circuit is zero. If the dipole loop moves to the left, a net current and force develop to push it back toward the right. Conversely, if it moves to the right, an opposite current and force push it back to the left.

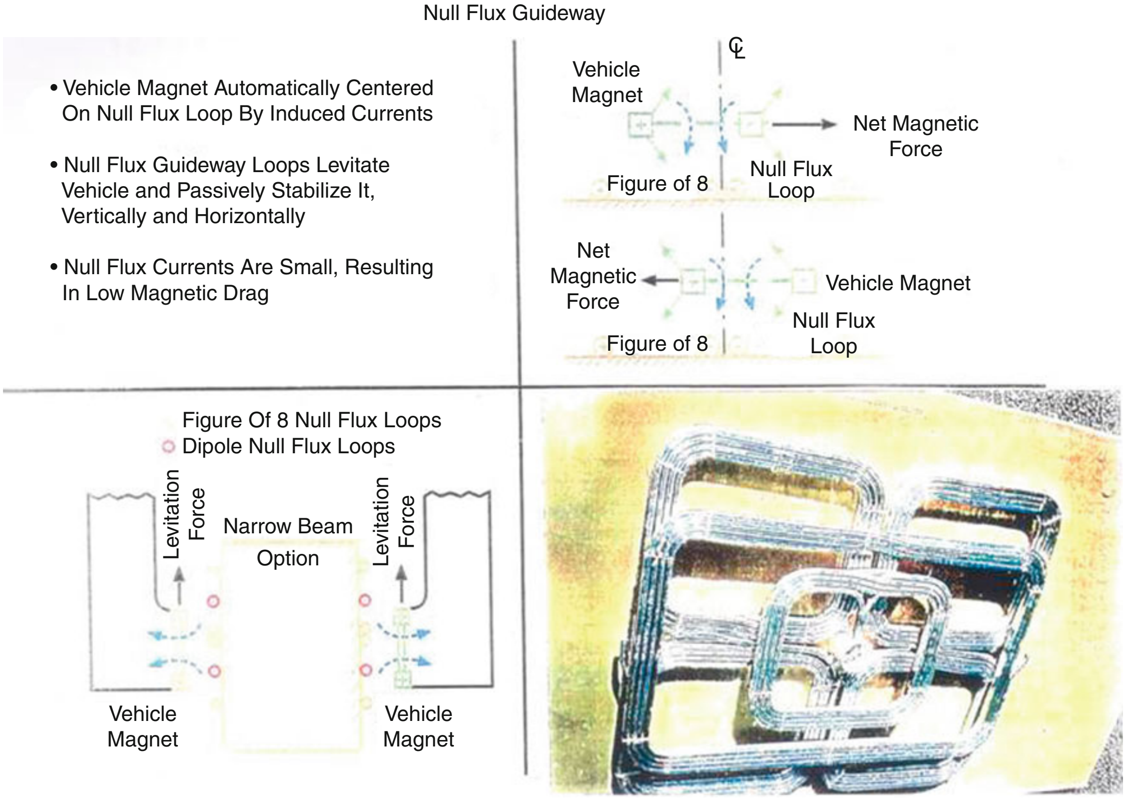

Superconducting Electrodynamic Maglev
Enabling Inventions

- Magnetic levitation using superconducting magnets and inductive normal guideway
- Null flux guideway
- Linear synchronous motor (LSM)
- Quadrupole vehicle magnets
- High speed electronic switches

MAGLEV Technology Development. Figure 20 Superconducting electrodynamic Maglev-enabling inventions

The lower left corner of Fig. 21 shows the guideway and vehicle loops for the narrow beam guideway. The figure of 8 loops, on the sides of the beam, levitate the vehicle. The vehicle magnets move slightly below the symmetry position, generating a current and force that push them upward. The figure of 8 null flux loops, besides levitating the vehicle, also stabilize it against vertical and roll displacement from external forces. The vehicle is stable against pitch displacements by having independent figure of 8 loops along the beam. If the vehicle tends to pitch downward due to an external force, the pitching motion is automatically countered by magnetic forces. A second null flux loop circuit on the narrow beam provides lateral and yaw stability. It consists of two dipole loops, one on each side of the beam, electrically connected to form a null flux circuit. When the vehicle is laterally centered on the beam, the net flux through the null flux circuit is zero. When the vehicle is laterally or yaw displaced from its centered position, a net flux and current develop in the dipole null flux circuit that pushes the vehicle back to the center.

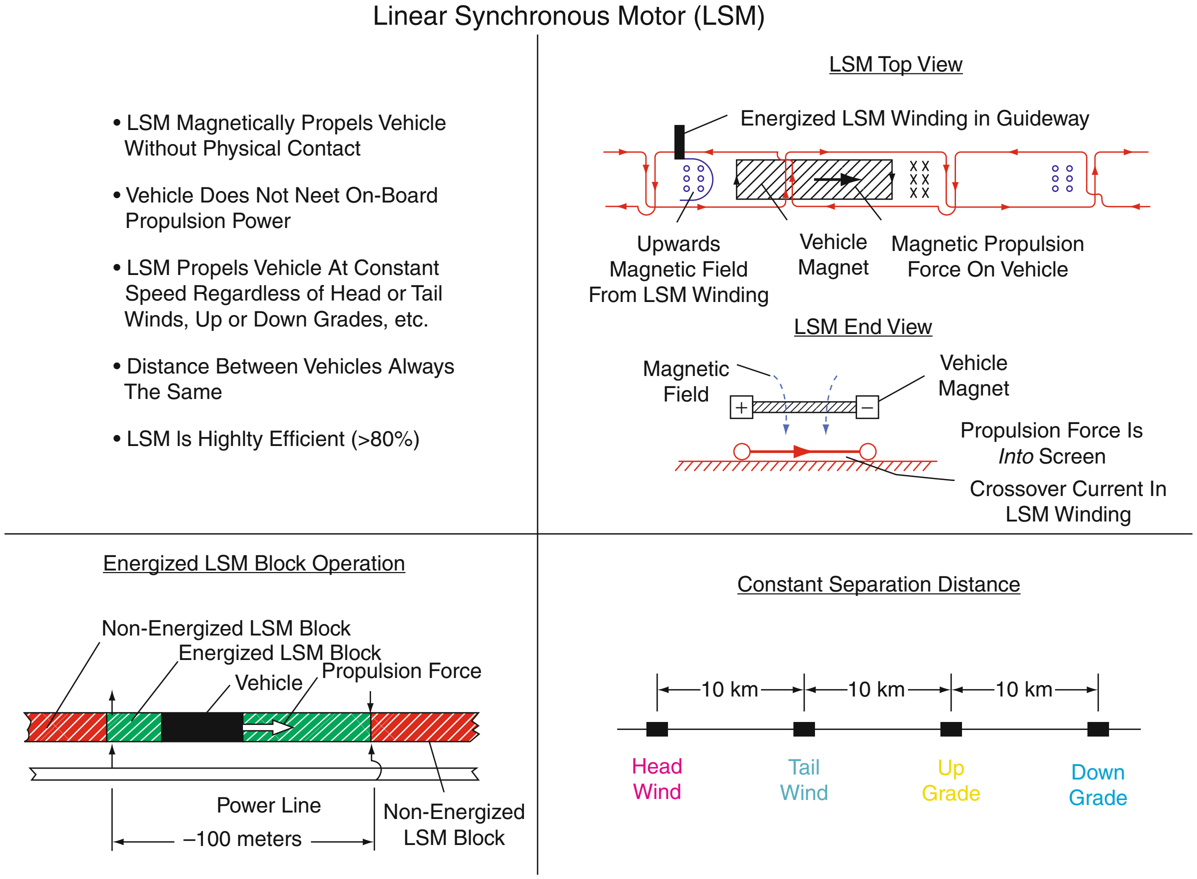
The lower right corner of Fig. 21 shows a fabricated guideway loop assembly for one side of the narrow beam. Besides the dipole and figure of 8 loops, there is also an LSM propulsion loop. The loop assembly is encapsulated in a thin panel of polymer concrete,


Null Flux Guideway
• Vehicle Magnet Automatically Centered On Null Flux Loop By Induced Currents
• Null Flux Guideway Loops Levitate Vehicle and Passively Stabilize It, Vertically and Horizontally
• Null Flux Currents Are Small, Resulting In Low Magnetic Drag
℄
Vehicle Magnet
Net Magnetic Force
Figure of 8
Null Flux Loop
Net Magnetic Force
Vehicle Magnet
Figure of 8
Null Flux Loop
Figure Of 8 Null Flux Loops
○ Dipole Null Flux Loops
Levitation Force
Narrow Beam Option
Levitation Force
Vehicle Magnet
Vehicle Magnet

MAGLEV Technology Development. Figure 21
Null flux guideway

a very strong, tough form of concrete, and attached to the sides of the narrow beam and shipped to the construction site. The I^2R losses in the guideway loops, and the corresponding magnetic drag, are much smaller in the null flux guideway than in a guideway that uses a simple conducting sheet or a set of simple dipole, non-null flux loops. Moreover, the magnetic "stiffness" – that is the magnitude of the magnetic restoring force per inch of displacement – is much greater for the null flux guideway.

The third invention is the linear synchronous motor (LSM) [7] illustrated in Fig. 22. Prior to the LSM, a method proposed for the propulsion of high speed ground transport vehicles included conventional steel wheel on rail traction motors, the linear induction motor (LIM), and jet engines. These methods required either on-board engines or collecting power from a high voltage line along the guideway using a sliding contact. In the linear synchronous motor, the superconducting magnets on the vehicle magnetically interact with a traveling wave of AC current in a set of aluminum conductor windings on the guideway, as illustrated in Fig. 22, producing a longitudinal magnetic force that propels the vehicle along the guideway. The vehicle remains locked in phase with the traveling AC current wave, with its speed determined by the frequency of the LSM current wave, in accordance with the relation $v = f\lambda$, where λ is the pitch of the alternating polarity magnets on the vehicle (λ = twice the length of one magnet), f is the frequency of the AC wave, and v is the vehicle speed. Vehicle speed is constant, even if it is subjected to head or tail winds or it climbs or descends grades, as long as sufficient phase margin is maintained by the LSM current wave. As a result, the distance between vehicles on the Maglev-2000 guideway is fixed, even if the external forces on the individual vehicles vary. This is extremely important since it eliminates any possibility of vehicle collisions.



Linear Synchronous Motor (LSM)

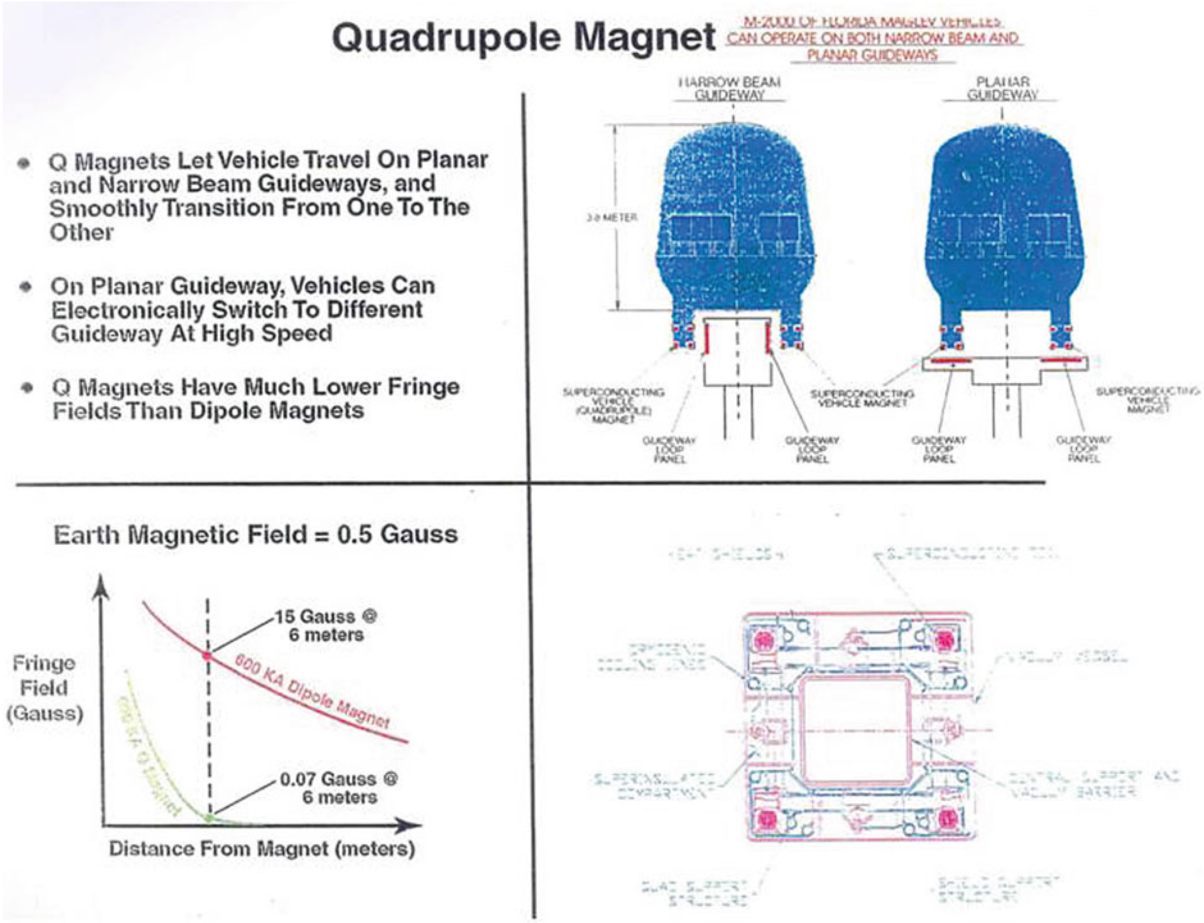
- LSM Magnetically Propels Vehicle Without Physical Contact
- Vehicle Does Not Neet On-Board Propulsion Power
- LSM Propels Vehicle At Constant Speed Regardless of Head or Tail Winds, Up or Down Grades, etc.
- Distance Between Vehicles Always The Same
- LSM Is Highlty Efficient (>80%)

LSM Top View
Energized LSM Winding in Guideway
Upwards Magnetic Field From LSM Winding
Vehicle Magnet
Magnetic Propulsion Force On Vehicle
LSM End View
Magnetic Field
Vehicle Magnet
+
−
Propulsion Force Is *Into* Screen
Crossover Current In LSM Winding
Energized LSM Block Operation
Non-Energized LSM Block
Energized LSM Block
Vehicle
Propulsion Force
Power Line
−100 meters
Non-Energized LSM Block
Constant Separation Distance
10 km
10 km
10 km
Head Wind
Tail Wind
Up Grade
Down Grade

MAGLEV Technology Development. Figure 22
Linear synchronous motor (LSM)

To minimize I^2R losses in the guideway LSM windings, the entire guideway is not energized. Instead, only the section on which the vehicle is traveling is switched on from the DC power line that runs along the guideway, using solid state switches to control local frequency and power flow. The switched-on section length is typically 100–200 m, resulting in high motor efficiency. Approximately 90% of the input electrical power to the switched section is delivered as propulsive power to the Maglev vehicle. As the vehicle leaves the old switched-on section, the electrical power input is transferred to the new section ahead, as illustrated in the drawing in the lower left corner of Fig. 22. By controlling the frequency of the AC power fed to the guideway, the central traffic control facility can speed up or slow down the vehicles if conditions warrant – again, a major advantage for safe operation.

The fourth invention is the Maglev-2000 quadrupole superconducting magnet. It differs from the dipole magnet used in the Japanese Maglev system, and achieves new capabilities. Dipole magnets employ two lines of current that run in opposite directions, forming a simple rectangular loop. Quadrupole magnets employ four lines of current, at the corners of a square, and alternating in direction as shown in Fig. 23. Fringe magnetic fields from quadrupole magnets are much smaller than from dipole magnets, as illustrated in the lower left corner of Fig. 23. A 600 KA dipole magnet has a fringe field of 15 gauss at 6 m, 30 times the Earth ambient field, while the quadrupole magnet fringe field is only 0.07 gauss, one-seventh of Earth ambient, at the same distance. The much smaller fringe fields make it simple to keep the magnetic field strength inside the vehicle passenger compartment and



Quadrupole Magnet

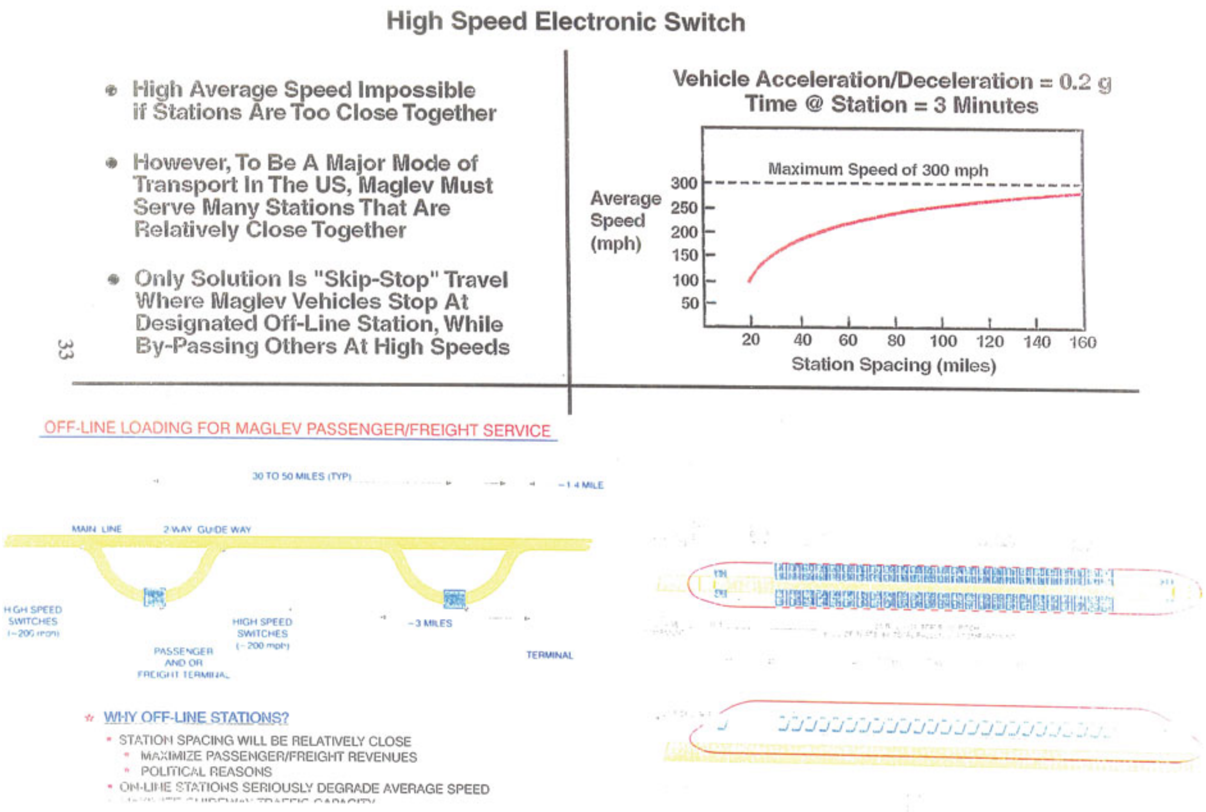
- Q Magnets Let Vehicle Travel On Planar and Narrow Beam Guideways, and Smoothly Transition From One To The Other
- On Planar Guideway, Vehicles Can Electronically Switch To Different Guideway At High Speed
- Q Magnets Have Much Lower Fringe Fields Than Dipole Magnets

M-2000 OF FLORIDA MAGLEV VEHICLES CAN OPERATE ON BOTH NARROW BEAM AND PLANAR GUIDEWAYS
NARROW BEAM GUIDEWAY
PLANAR GUIDEWAY
3.0 METER
SUPERCONDUCTING VEHICLE (QUADRUPOLE) MAGNET
SUPERCONDUCTING VEHICLE MAGNET
SUPERCONDUCTING VEHICLE MAGNET
GUIDEWAY LOOP PANEL
GUIDEWAY LOOP PANEL
GUIDEWAY LOOP PANEL
GUIDEWAY LOOP PANEL
Earth Magnetic Field = 0.5 Gauss
Fringe Field (Gauss)
15 Gauss @ 6 meters
600 KA Dipole Magnet
600 KA Q Magnet
0.07 Gauss @ 6 meters
Distance From Magnet (meters)
SUPERINSULATED COMPARTMENT
CENTRAL SUPPORT AND VACUUM BARRIER

MAGLEV Technology Development. Figure 23
Quadrupole magnet

at stations at Earth ambient level. Quadrupole magnets enable Maglev vehicles to travel on and smoothly transition between both narrow beam and planar guideways. Most of the time, Maglev-2000 vehicles travel on the low cost narrow beam guideway. Near stations, the vehicles transition to the planar guideway where the vehicle can switch from the main guideway to off-line stations to stop and carry out loading/unloading operations. The switch section has two lines of overlapping planar guideway loops, with one line going on as the main high speed route, and the second line directing the vehicle to an off-line station away from the main route. The switching process uses electronic solid state switches to control which line of guideway loops the vehicle will follow. Mechanical movement of the guideway is not required. Depending on schedule, the vehicle may not switch off the guideway, but continue at high speed on the main line to a station further down the line.

The capability to switch at high speed to off-line stations for loading and unloading using an electronic, nonmechanical switching process – the fifth fundamental invention – is very important for countries like the USA, where the population density is relatively low and spread out, not concentrated at a few high density points. High speed switching enables Maglev vehicles to employ "skip-stop" schedules to maintain high average speed, as illustrated in Fig. 24. If vehicles had to stop at or slow down for every station along the route, the average speed would be considerably less than 300 mph, unless stations were a hundred miles apart or more. With "skip-stop" scheduling, stations can be spaced more



High Speed Electronic Switch

- High Average Speed Impossible If Stations Are Too Close Together
- However, To Be A Major Mode of Transport In The US, Maglev Must Serve Many Stations That Are Relatively Close Together
- Only Solution Is "Skip-Stop" Travel Where Maglev Vehicles Stop At Designated Off-Line Station, While By-Passing Others At High Speeds

33

OFF-LINE LOADING FOR MAGLEV PASSENGER/FREIGHT SERVICE

30 TO 50 MILES (TYP)
~1.4 MILE
MAIN LINE
2 WAY GUIDE WAY
HIGH SPEED SWITCHES (~200 mph)
HIGH SPEED SWITCHES (~200 mph)
~3 MILES
PASSENGER AND OR FREIGHT TERMINAL
TERMINAL

WHY OFF-LINE STATIONS?

- STATION SPACING WILL BE RELATIVELY CLOSE
  - MAXIMIZE PASSENGER/FREIGHT REVENUES
  - POLITICAL REASONS
- ON-LINE STATIONS SERIOUSLY DEGRADE AVERAGE SPEED

MAGLEV Technology Development. Figure 24
High speed electronic switch

closely to more efficiently serve the spread-out population. Using "skip-stop" travel, a passenger could board a Maglev vehicle at his/her local station and skip all other stations in the area, traveling directly to his/her final destination in some distant city.

Operational First-Generation Maglev Systems

Following the publication of Powell and Danby's 1966 paper on superconducting Maglev [8], a number of countries started Maglev development programs. Japan, and also Germany, visited Powell and Danby to learn about their inventions. Japan chose to focus their Maglev development program on superconducting Maglev because of its large clearance of 10 cm or more – a very important factor in Japan because of the strong earthquakes there – and its strong inherent stability, a very important safety feature.

Germany's Maglev development program focused on both superconducting electrodynamic Maglev and electromagnetic Maglev systems, and tested both approaches. They then selected electromagnetic Maglev as a system that could be commercialized relatively quickly.

Japan and Germany have both achieved technically successful first-generation Maglev systems that have transported many thousands of passengers safely and comfortably at speeds on the order of 300 mph. The Japanese Maglev system presently operates on Japan Railways' 20 km demonstration track in Yamanashi Prefecture [17]. Japan Railways plans to have the Yamanashi track become part of a 500 km (300 mile) Maglev route to be constructed between Tokyo and Osaka that will carry 100,000 passengers daily with a trip time of 1 h [18].

The German electromagnetic Maglev system, called Transrapid, was extensively tested on a 20 km demonstration track at Emsland, Germany [5]. The Transrapid Maglev system now operating in Shanghai, China [19], has the distinction of being the first commercial Maglev system. The 30 km route runs between the center of Shanghai and its Pudong airport.

Both the Japanese and German first-generation Maglev systems are limited to transporting passengers and lightweight, high-value FedEx-type freight. They cannot transport heavy trucks, autos, and freight containers. They cannot electronically switch from the main guideway to a secondary guideway, but require mechanically moving long sections of the guideway. Guideway construction cost is very high, on the order of $60 million per two-way mile or more [20].

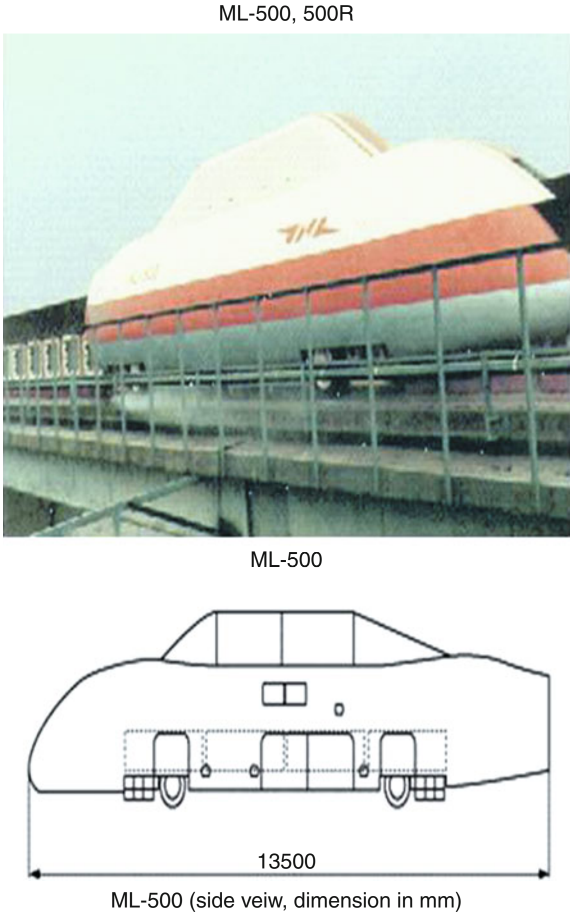
More detailed descriptions of the development and current status of the Japanese and German first-generation passenger Maglev system is given below.

Table 1 [21] lists the significant milestones in the development of the Japanese first-generation Maglev system from its start in 1970 through 2004, the last available update from Japan Railways' Railway Technical Research Institute (RTRI). By 2004, the Yamanashi Maglev Demonstration Line had carried over 80,000 passengers [21] with total accumulated running distances of over 400,000 km (240,000 miles). Since then, the number of passengers and running miles has grown considerably, but data is not available.

The 20 km Yamanashi line located north of Mt. Fuji was dedicated in April 1997. Before then, R&D tests were carried out at the Miyazaki Test site. The first test run of a levitated Maglev vehicle was in 1977. The unmanned ML-500 vehicle (Fig. 25) ran on an inverted T-shaped guideway. In 1977, it attained a speed of 517 km/h [21].

MAGLEV Technology Development. Table 1 Milestones in the development of the first-generation Japan Railways' superconducting Maglev system [21]

Date	Milestone
1970	Study of electrodynamic levitation system (EDS) started
1977	Miyazaki Maglev Test Center opened – test run of unmanned ML-500 vehicle on inverted T-shaped guideway started
1980	Test run of MLU 001 vehicle on U-shaped guideway started at Miyazaki
1982	Manned 2 MLU 001 vehicle set tests started
1986	3 MLU 001 vehicle set achieved 352 km/h
1987	400 km/h achieved by manned MLU 002 vehicle set
1997	First running test of MLX 01 vehicle on Yamanashi Demonstration Line Manned 5 MLX 01 vehicle set attained 552 km/h
1999	Test of 2 MLX 01 vehicle sets passing each other in opposite directions at 1,003 km/h
2003	Manned 3 MLX 01 vehicle set attained 581 km/h (361 mph)
2004	Cumulative number of passengers carried at Yamanashi exceeds 80,000; cumulative distance traveled exceeds 400,000 km

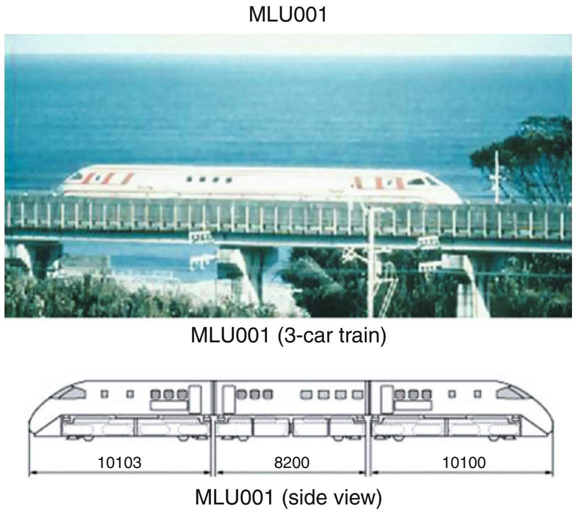


MAGLEV Technology Development. Figure 25 ML-500, 500R

Japan Railways then transitioned to a U-shaped guideway, which is the configuration for the Yamanashi demonstration route, and the planned line between Tokyo and Osaka. The first test run of the MLU 001 (Fig. 26) vehicle [21] on the U-shaped guideway was carried out in 1980 at the Miyazaki Test site. The U-shaped guideway permitted a more efficient configuration for carrying passengers. Testing of the MLU001 vehicle carrying passengers started in 1982, with a two-car train set. In 1987, the two-car manned vehicle set attained 400 km/h. Also in 1987, test runs of a new Maglev vehicle, MLU002 (Fig. 27), began [21]. It was longer than the MLU001 – 22 m compared to 20 m for the two-car MLUJ001 vehicle set. MLU002 could carry 44 passengers and weighed 17 t. In 1989, it attained a speed of 394 km/h.

In 1991, MLU002 was damaged by a fire started in its braking system caused by excessive heating. MLU002 was then replaced by MLU002N (Fig. 28) which was essentially the same length as MLU001 and slightly heavier, 19 t compared to 17 t [21]. Test runs of MLU002 started in 1993. In 1995, MLU002 attained a speed of 411 km/h with passengers.

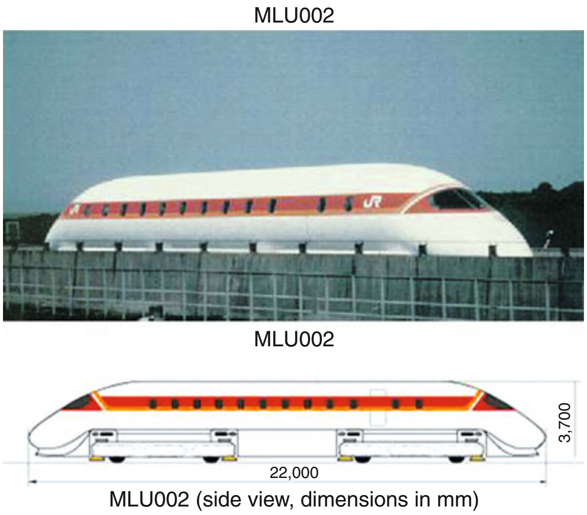
By 1996, the Yamanashi Test Center opened, and was officially dedicated in April 1997. A few months later, in December, the new MLX01 vehicle (Fig. 29)


MLU001

MLU001 (3-car train)

MLU001 (side view)

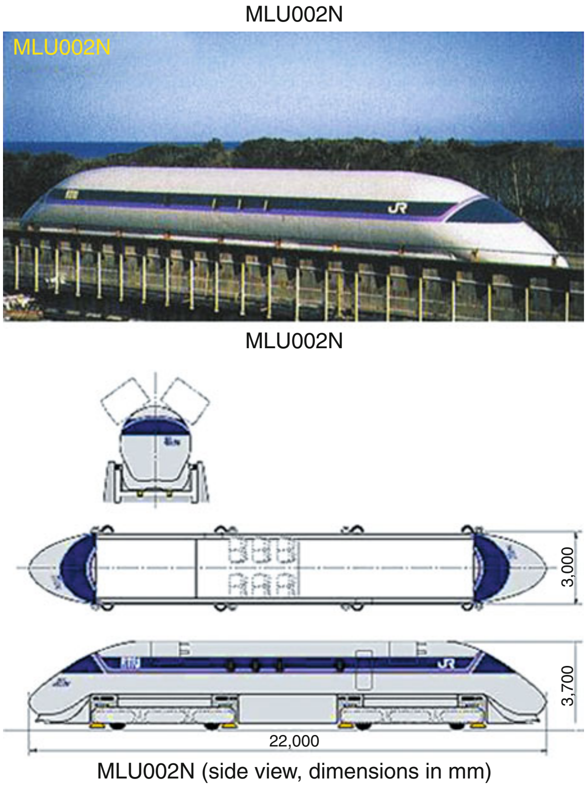
MAGLEV Technology Development. Figure 26
MLU001


MLU002

MLU002

MLU002 (side view, dimensions in mm)

MAGLEV Technology Development. Figure 27
MLU002


MLU002N

MLU002N

MLU002N (side view, dimensions in mm)

MAGLEV Technology Development. Figure 28
MLU002N

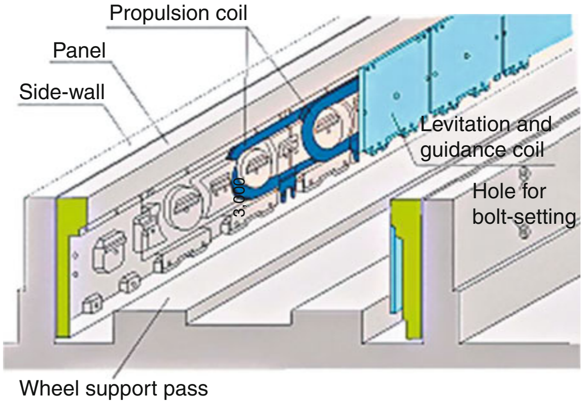
MLX01 leading car, aero-wedge

MLX01 leading car, double cusp

MAGLEV Technology Development. Figure 29
MLX01

New leading car, MLX01-901

MAGLEV Technology Development. Figure 30
MLX01-901



MAGLEV Technology Development. Figure 31
Panel method

attained 531 km/h carrying passengers. In 1999, a five-vehicle MLX01 set reached 552 km/h with passengers [21] and two-vehicle sets passed each other going in opposite directions with a relative speed of 1003 km/h (600 mph). Tests on a new vehicle set, MLX01-901 (Fig. 30) started in 2002. A three-vehicle MLX-01 set attained the world speed record of 581 km/h (361 mph) on December 2, 2003, carrying passengers [21].

The Japanese Maglev system uses a U-shaped guideway, with the aluminum levitation, stability, and propulsion located on the beams at the sides of the guideway (Fig. 31). The concrete panels with their attached aluminum loops are manufactured in a factory and then transported to the construction site [21] where they are bolted to the concrete side beams of the massive field constructed U-shaped guideway. Figure 32 shows a photo of the completed guideway.

Figure 33 shows a photo of the Yamanashi Maglev Demonstration line, with two Maglev vehicles on a bridge that crosses a local highway. In the distance, the JR Maglev guideway enters a tunnel through the mountain. Much of the proposed 500 km (300 miles) Maglev route will be in deep tunnels through the mountains of Central Japan. Three routes are under consideration [22] as illustrated in Fig. 34. Depending on the route, as much as 60% of the Tokyo to Osaka line will be in deep tunnels.

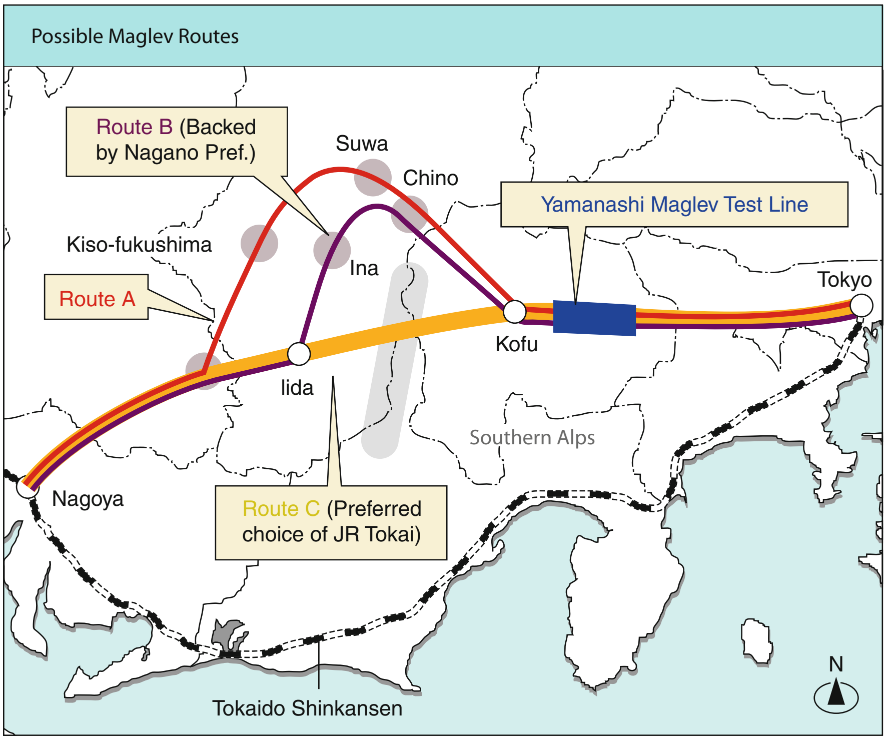
The principal constraints for the first-generation Japanese Maglev system are its high construction cost and its limitation to passenger-only transport. The high field construction cost results from the massive U-shaped guideway, plus the high cost of tunneling for much of the Tokyo to Osaka route. The limitation to passenger-only transport can be a revenue problem in low population countries, like the USA. In the USA, for example, the transport outlay for intercity trucks is much greater than for intercity air and train passengers – the USA spends over $300 billion per year for intercity truck transport, compared to only

MAGLEV Technology Development. Figure 32
Guideway of the Yamanashi Maglev test line

MAGLEV Technology Development. Figure 33
Vehicles crossing bridge over highway on Yamanashi Maglev test line

60 billion for air passengers and three billion for intercity rail passengers.

Turning now to the German electromagnetic Transrapid Maglev system, it is the basis for the first high speed commercial Maglev system in the world. Transrapid is a first-generation electromagnetic (EMS) Maglev system [5] which is levitated by conventional electromagnets on the vehicle that are attracted upward to two iron rails on the guideway. The physical gap between the vehicle's electromagnets is small,



MAGLEV Technology Development. Figure 34
Possible Maglev routes

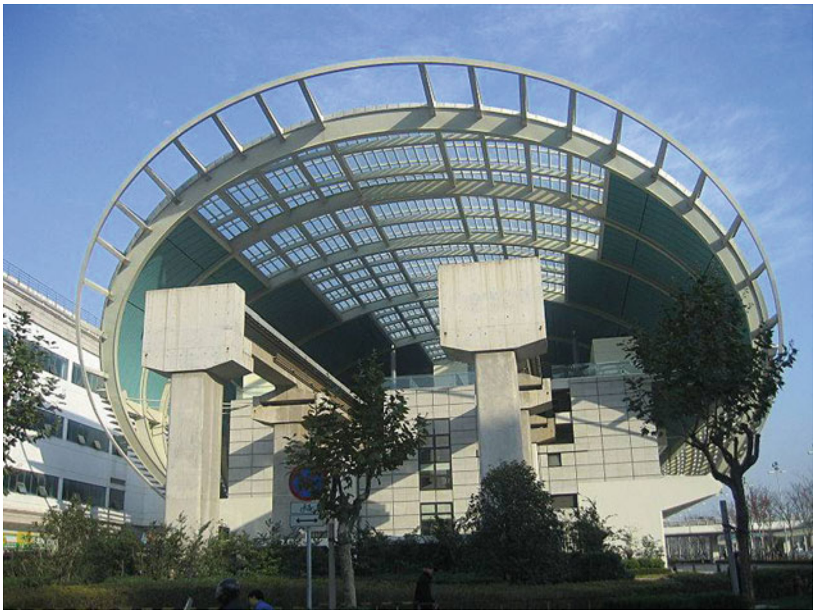
approximately 1 cm (3/8 in.), necessitating very precise and accurate guideway construction. EMS suspensions are inherently unstable because as the gap between the vehicle's electromagnets and the iron rail decreases, the levitating force on the vehicle increases, acting to pull the vehicle into contact with the rail. The Transrapid system rapidly servo controls the currents in the electromagnets on the vehicle to keep the gap between the vehicle and the guideway constant. If the vehicle moves toward the iron rails above the electromagnets, the servo control system quickly decreases the current in their windings, reducing the levitating force. If the vehicle moves down and away from the iron rails, the current in the electromagnets is quickly raised, increasing the levitating force. The servo control system operates on a time scale of thousandths of a second.

The 30 km long Shanghai Maglev Line runs from the outskirts of Central Shanghai to Pudong Airport. Figure 35 shows a photo of the Transrapid vehicle leaving Pudong Airport, while Fig. 36 shows a photo of the Longyang Road Station at the outskirts of Shanghai. The Transrapid Maglev Line does not service the center of Shanghai. Instead, there is a 20 min subway ride from the center of Shanghai to the Longyang Road Maglev Station [19].

Maglev service on the Shanghai route started on January 1, 2004. Depending on time of day, trip time for the 30 km run is 7.20 min for peak travel hours and 8.10 min for nonpeak hours. The corresponding average speeds are 251 km/h (156 mph) for the faster trip and 224 km/h (139 mph) for the slower trip. Maximum speeds during the trip are 431 km/h (268 mph) and 301 km/h (197 mph). The record speed on the route was 501 km/h (311 mph), achieved on November 12, 2003. The interval between trains is 15 min, except in the evening, when it is 20 min [19].

The 30 km (19 miles) Shanghai Transrapid line cost $1.8 billion (US dollars) and was built in 2½ years. The construction cost was $60 million per kilometer of route ($95 million per mile) [19]. The proposed 200 km (124 miles) extension to Hangzhou was approved in March 2010, with construction planned to start in 2010 [19]. However, given the high cost per kilometer

MAGLEV Technology Development. Figure 35
Maglev train coming out of the Pudong International Airport in Shanghai

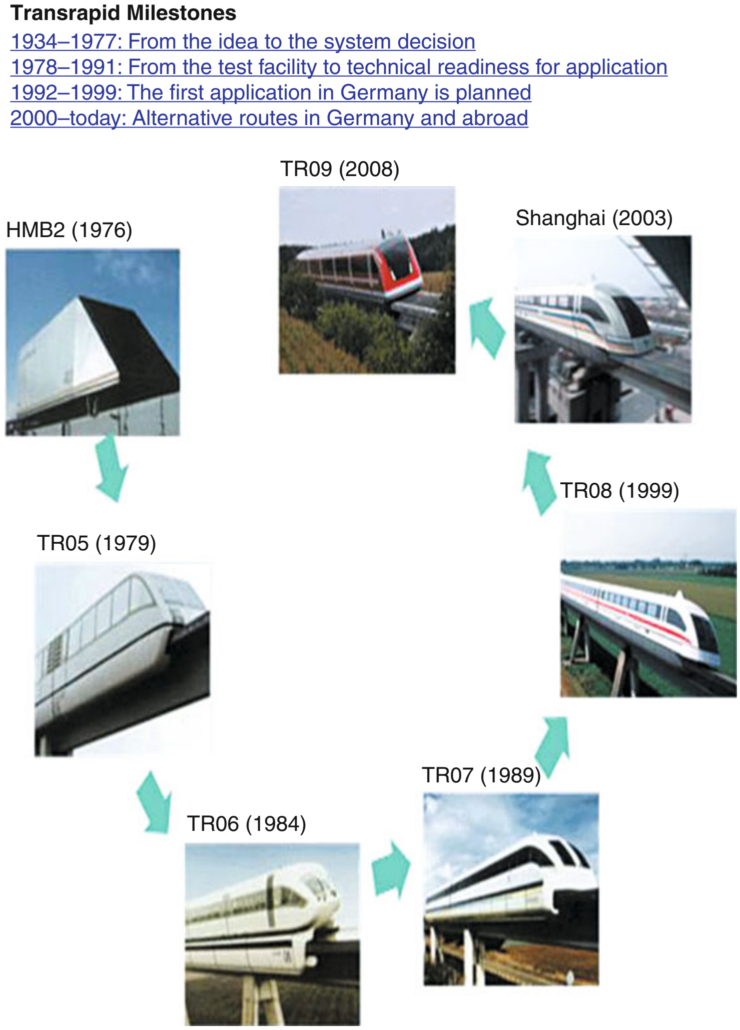


MAGLEV Technology Development. Figure 36
Longyang road Maglev station

and China's plans to build thousands of kilometers of high speed rail, the proposed Maglev extension may be postponed or even canceled.

The development of the Transrapid Maglev System has proceeded through a number of vehicles, as illustrated in Fig. 37, eight during the years 1976–2008. A number of EMS vehicles were tested in Germany in prior years, before 1976, by various companies before settling the final Transrapid system. These included the MBB "Prinzipfahrzeng," the MBB Komet, and the Krauss-Maffei Transrapid TR-02 and TR04 vehicles. In 1974, MBB and Krauss-Maffei joined together to form Transrapid.

The speed records achieved by the various Transrapid vehicles and their predecessors are given below [23]:



Transrapid Milestones

1934–1977: From the idea to the system decision
1978–1991: From the test facility to technical readiness for application
1992–1999: The first application in Germany is planned
2000–today: Alternative routes in Germany and abroad

MAGLEV Technology Development. Figure 37
Transrapid milestones. 1934–1977, from the idea to the system decision; 1978–1991, from the test facility to technical readiness for application; 1992–1999, the first application in Germany is planned; 2000–present, alternative routes in Germany and abroad

Year	Vehicle	Speed, km/h
1971	Rinzipfahrzeg	90
1972	TR02	164
1973	TR04	250 (manned)
1975	Komet	401
1987	TR06	406 (manned)
1988	TR06	412 (manned)
1989	TR07	436 (manned)
1993	TR07	450 (manned)
2003	TR08	501 (manned)

For extended testing of the Transrapid vehicles, Germany constructed a 31.5 km test track at Emsland in Germany. The single track had turning loops at both ends, permitting continuous running tests. Construction of the Emsland Facility began in 1980 and was completed in 1984. It was the site of the 2006 accident when a moving Transrapid vehicle collided with a stopped maintenance vehicle that was working on the guideway [5].

The Transrapid vehicles through TR08 obtained power for levitation and electronics below 80 km/h by physical contact with the track. The new TR09 vehicle (Fig. 38) needs no physical contact with the track at any speed, but receives power by inductive transmission with the track's propulsion system. If the track's propulsion power system fails, TR09 uses on-board batteries for back power to maintain levitation [5].

Application of Transrapid in Germany has been prevented by two factors: (1) the existing network of ICE high speed rail lines, and (2) high construction cost. The proposed Hamburg to Berlin line was canceled because of cost, as was the Munich to its airport route. The 40 km (25 mile) route between Munich Central Station and Munich Airport was originally estimated to cost 1.85 € billion, but later raised to over 3 € billion ($4.7 billion) or almost $200 million per mile. High tunnel and civil engineering costs were significant factors in the 2008 decision to cancel the project [5].

Transrapid has been intensively lobbied in the USA to build Maglev routes between major metropolitan areas, for example, Las Vegas to Anaheim, California, Baltimore to Washington, DC, Pittsburgh to its airport, etc. So far, however, no projects have gone forward.

The Development of the Second-Generation Maglev-2000 System

The first-generation Maglev systems, while technically successful, have two factors that limit their implementation, particularly in the USA.

First, the guideway construction cost is very high, $60 million or more per mile. Second, the first-generation Maglev systems only carry passengers. While useful in densely populated Europe and Japan, passenger-only systems are less useful in lower population density large countries like the USA. US transport outlays for intercity trucks over $300 billion annually; for intercity air passengers, $60 billion per year; and for intercity rail passengers, only $3 billion per year [24]. Maglev or high speed rail passengers only systems in the USA will require major government financing for construction and large subsidies for operation and maintenance. Because of USA's large debt, major government financing is unlikely. Maglev routes will need private investment. To achieve this, they must be profitable, with a short payback time on invested capital, less than a decade.

MAGLEV Technology Development. Figure 38
Transrapid 09 at the Emsland test facility in Germany

The second-generation Maglev-2000 system addresses these two factors. First, the projected guideway construction cost is about $25 million per mile, a factor of 2 or more lower than first-generation systems. Low cost prefabricated monorails are used for the elevated guideway. Figure 39 shows an artist's view of a Maglev-2000 passenger vehicle on the monorail guideway.

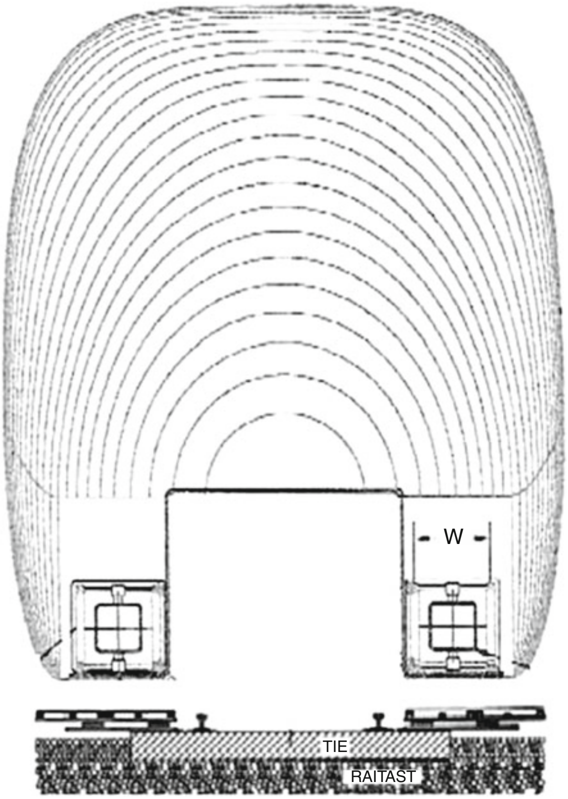
The prefabricated monorail beams would be mass produced in factories, with their guideway loop panels, sensors, electronic equipment, etc., attached to them at the factory. The beams and piers would then be transported by truck or rail to the construction site, where they would be quickly erected on pre-poured concrete footings or pilings, using conventional cranes. Low guideway cost would be achieved by the use of conventional box beams for the monorail, which minimizes the materials required, and prefabrication, which minimizes expensive field construction. Disruptions to local infrastructure would also be minimized, which would reduce local opposition to guideway construction.

The superconducting quadrupoles on the vehicles (Fig. 23) have four magnetic poles that alternate in polarity around their circumference. When on the monorail guideway, the vertical sides of the quadrupoles interact with the aluminum loops attached to the adjacent vertical sides of the monorail guideway. When operating on a planar guideway (Fig. 23), the bottom faces of the quadrupoles magnetically interact with aluminum loops located on the guideway beneath the vehicle.

The ability to operate on a planar guideway as well as monorail reduces construction cost. When operating in densely populated urban and suburban areas, Maglev-2000 vehicles do not need a new, very expensive guideway with its accompanying disruptions to existing infrastructure. Instead, the vehicles can transition to, and operate on, existing RR tracks to which aluminum loop guideway panels have been attached on the cross-ties (Fig. 40). Conventional trains can continue to use the RR tracks, given appropriate scheduling. The cost of attaching guideway panels to enable levitated Maglev-2000 operation is very small, only about $4 million per mile, compared to the much higher cost of building a new elevated guideway.

MAGLEV 2000

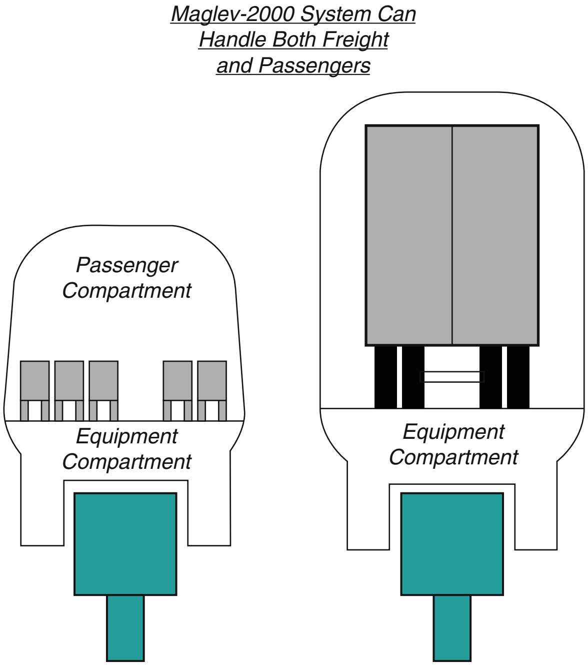
MAGLEV Technology Development. Figure 39
Artist's drawing of Maglev-2000 vehicle on monorail guideway



MAGLEV Technology Development. Figure 40
Drawing of levitated Maglev-2000 vehicle traveling on conventional RR tracks to which aluminum loop panels have been attached to the cross-ties

Turning to the second factor affecting Maglev implementation, revenues and net profits, Maglev-2000 can carry heavy trucks as well as passenger vehicles on its dual-use guideway. In contrast to the superconducting dipole loops used in the Japanese first-generation Maglev system, the powerful superconducting Maglev-2000 quadrupoles can be located along the length of a Maglev vehicle, increasing its lifting capability without producing magnetic fields inside the vehicle that significantly exceed the natural Earth ambient value. This is a result of the considerably lower value of the magnetic fringe fields from a quadrupole, compared to a dipole system.

Figure 41 shows a drawing of a passenger and truck carrier Maglev-2000 vehicles on the dual-use Maglev-2000 monorail guideway. The gross revenue from transporting 3,000 trucks daily on a Maglev-2000 route (one-fifth of the daily truck traffic on a typical Interstate Highway) is equivalent to 180,000 passengers per day, assuming a typical US outlay of 30 cents a ton mile for trucks, and 10 cents a passenger mile. The truck revenues alone would pay back the Maglev-2000 guideway in less than 5 years, attractive for private investment.



MAGLEV Technology Development. Figure 41
Maglev-2000 passenger and truck carrier vehicles on dual-use guideway

The same Maglev-2000 guideway could also transport personal autos together with their passengers, on long trips at lower cost than by highway. It could also transport high-value freight.

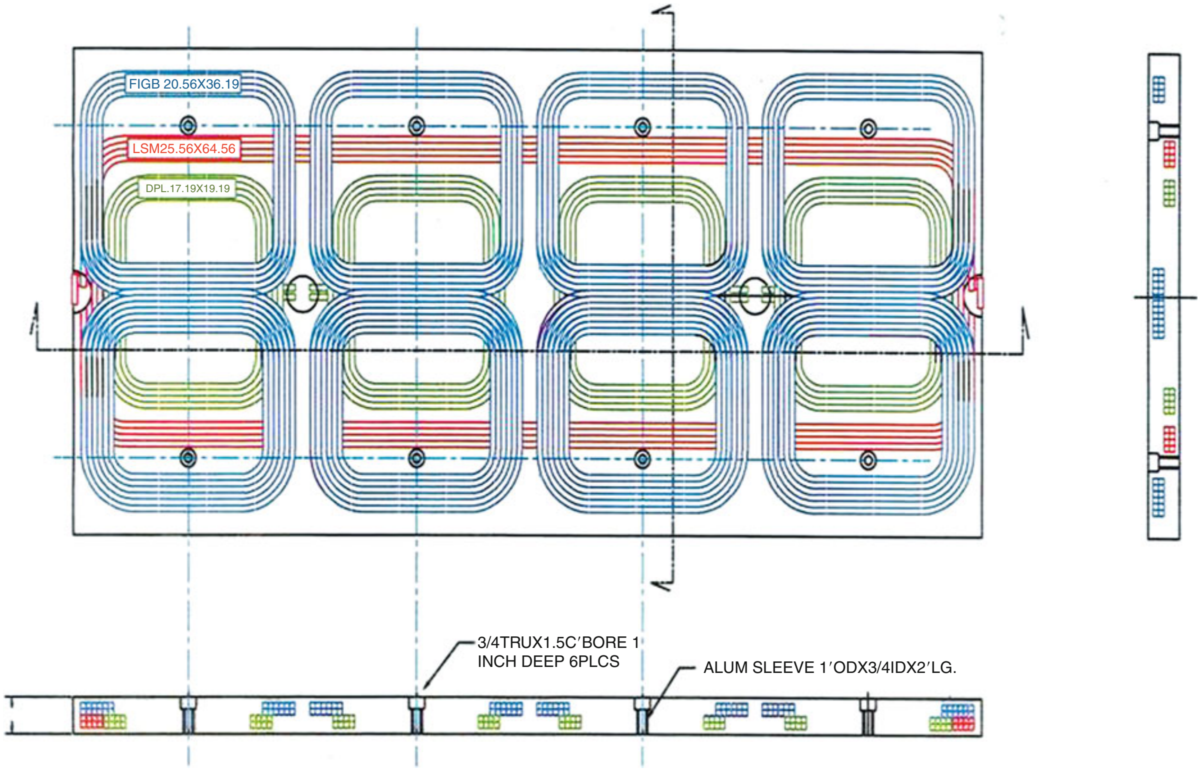
Using Maglev, a trucker could pick up a load, drive a few miles to the nearest Maglev station, travel across the USA in a few hours rather than a few days, and drive off the Maglev vehicle to deliver the load to its destination a few miles away. Wear and tear on the truck would be much less and it could make five deliveries in the time it would take to go by highway.

The ability of Maglev-2000 vehicles to travel on planar guideways also enables high speed electronic switching to off-line stations. Maglev-2000 vehicles can bypass high speed stations they are not scheduled to stop at, enabling stations to be closely spaced for convenient access. This increases revenue potential compared to only a few stations in a metropolitan area.

Figure 42 shows a drawing of the Maglev-2000 aluminum wire loop guideway panel. It has three sets of multi-turn aluminum loops: (1) a sequence of four short figure of 8 loops; (2) a sequence of four short dipole loops; and (3) one long dipole loop.

When the panels are attached on the vertical sides of the monorail guideway beam, the figure of 8 loops provide levitation and vertical stability. The dipole loops on each side of the beam are connected together to make a null flux circuit that maintains the vehicle in a centered position on the beam – when centered no current flows in the aluminum null flux circuit. When an external force (wind, curve, etc.) acts to push the vehicle away from its centered position, a magnetic force develops that opposes the external force. The long dipole loop is part of the linear synchronous motor (LSM) propulsion system, in which the loops on a sequence of panels are connected in series to form an energized block along which the Maglev vehicles travel. The energized block is typically on the order of 100 m in length; as the vehicle leaves an energized block, its AC propulsion current is switched into the next block that the vehicle is entering.

For the planar guideway, the panels are laid flat on the planar surface beneath the line of quadrupoles on the moving vehicle. The figure of 8 loops now provide lateral stability, generating magnetic restoring forces if


FIGB 20.56X36.19
LSM25.56X64.56
DPL.17.19X19.19
3/4TRUX1.5C′BORE 1 INCH DEEP 6PLCS
ALUM SLEEVE 1′ODX3/4IDX2′LG.

MAGLEV Technology Development. Figure 42
Drawing of aluminum loop guideway panel providing vertical life and stability, lateral stability, and linear synchronous propulsion

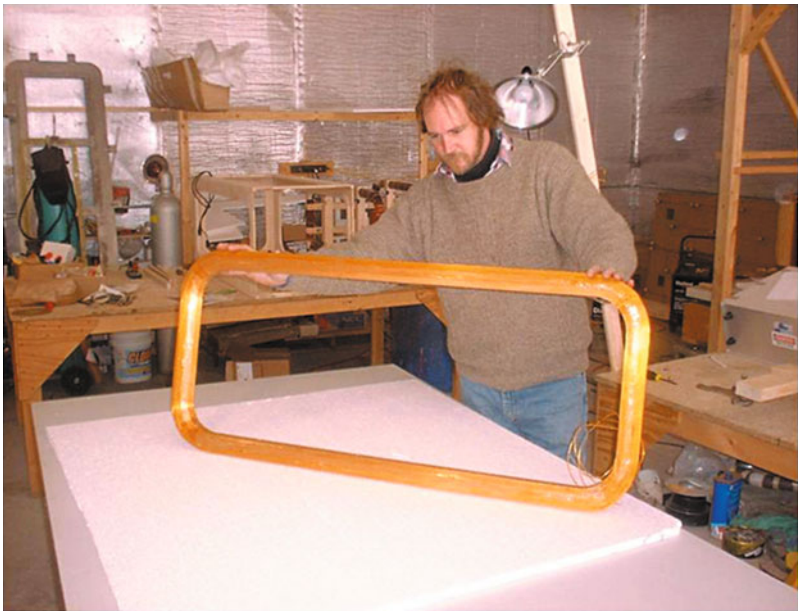
an external force acts to displace the vehicle from its centered position on the guideway. The dipole loops act individually, with inductive currents that levitate overhead. The LSM loops function as they do on the monorail guideway.

The planar guideway panel configuration can levitate and propel Maglev vehicles along existing RR tracks, with the panels attached to the cross-ties of the RR tracks.

The planar guideway also enables high speed switching. At a switch section, there are two lines of overlapping guideway loops, which can be either electronically open-circuited or close-circuited, depending on the desired switching action. Line A of guideway loops runs straight ahead on the main line, while the second line (Line B) of loops diverges laterally at a rate acceptable to passengers. If the loops in Line A are close-circuited and those in Line B are open-circuited, the vehicle travels straight ahead on the main route. If Line A is open and Line B is closed, the vehicle diverges laterally from the main route onto a secondary guideway leading to the off-line station. The high speed vehicle then decelerates on the secondary guideway that leads into the off-line station. When the vehicle leaves the station to rejoin the main line, it accelerates on an out-bound secondary guideway to a second switch section where the high speed vehicle reenters the main line.

The Maglev-2000 components discussed above have been fabricated and tested at full scale in order to determine their performance and validate their projected costs.

Figure 43 shows one of the two wound superconducting loops used for the Maglev-2000 quadrupole. The loop has 600 turns of NbTi superconducting wire, supplied by Supercon, Inc. of Shrewsbury, MA [25]. At the design current of 1000 A in the NbTi wire, the Maglev-2000 quadrupole has a total of 600,000 A turns in each of its two superconducting (SC) loops. The SC winding is



MAGLEV Technology Development. Figure 43
NbTi superconductor loop for Maglev-2000 quadrupole

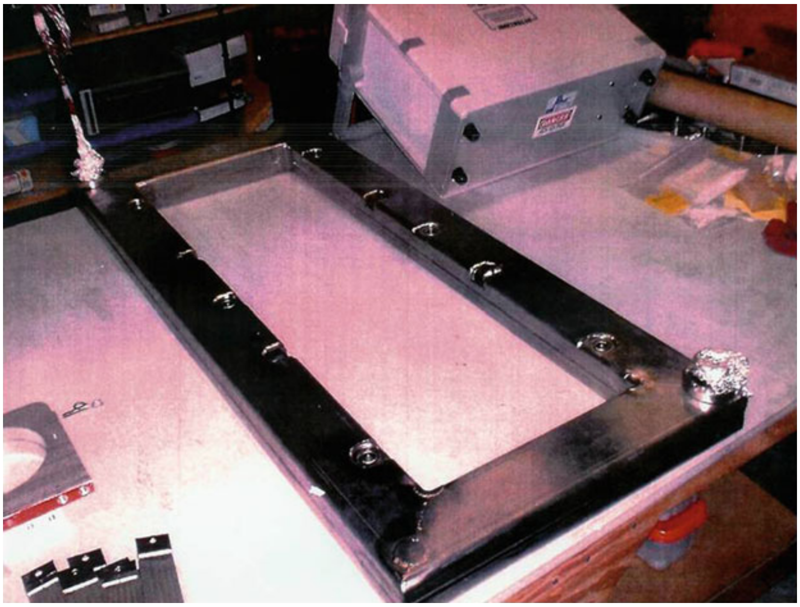
porous, with small gaps between the NbTi wires to allow liquid helium flow to maintain their temperature at 4.2 K, and to stabilize them against flux jumps and micro movements [25].

Figure 44 shows the SC loop enclosed in its stainless steel jacket. Liquid helium flows into the jacket at one end and exits at the end diagonally across from the entrance, providing continuous helium flow through the SC winding. Before insertion of the SC loop into the jacket, it is wrapped with a thin sheet of high purity aluminum (5000 residual resistance ratio) to shield the NbTi superconductor from external magnetic field fluctuations. After closing the jacket, a second layer of high purity aluminum is wrapped around it for additional shielding.

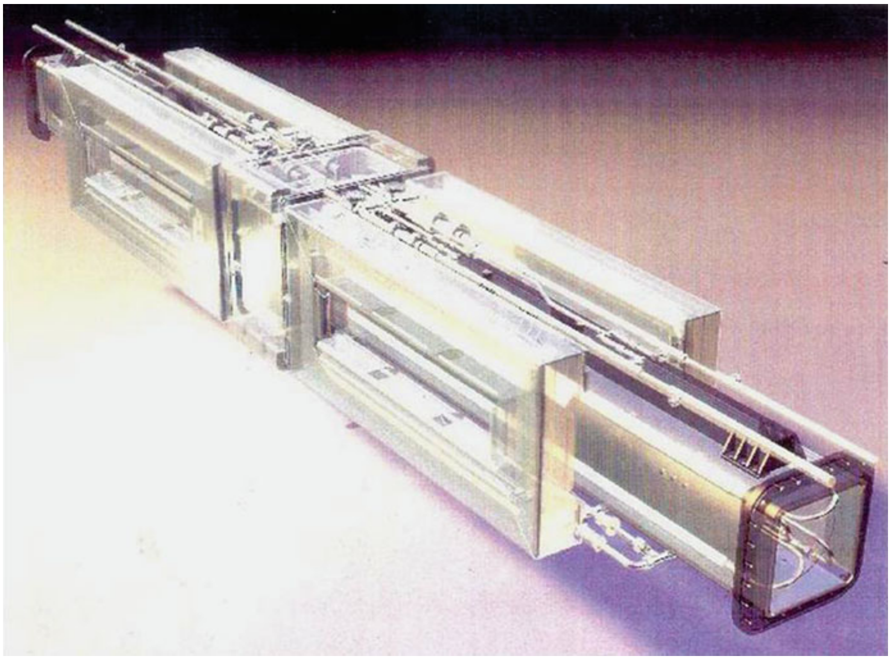
Figure 45 shows a CAD-CAM drawing of the complete Maglev-2000 cryostat that holds two superconducting quadrupoles. The magnetic polarity of the front SC quadrupole is opposite to that of the rear quadrupole. This allows levitation at lower speed than if the two quadrupoles had the same polarity, due to less L/R decay of the currents induced in the aluminum guideway loops. The two SC loops are supported by a graphite-epoxy composite structure that resists the magnetic forces – due both to the forces in a loop from its self current, and to the forces between the two loops – that act on them.

Figure 46 shows the SC loops, support structure, and cooling currents for the Maglev-2000 quadrupole being assembled in Maglev-2000's facility on Long Island. The SC loops have a 10 K thermal shield, which is cooled by helium exiting from the jacket holding the SC loop. The SC quadrupole structure is then enclosed by an outer layer of multilayer insulation (MLI) consisting of multiple alternating layers of glass fiber and aluminum foil. A second thermal shield encloses the SC quad, and maintained at ~70 K by the helium outflow from the 10 K primary thermal shield.

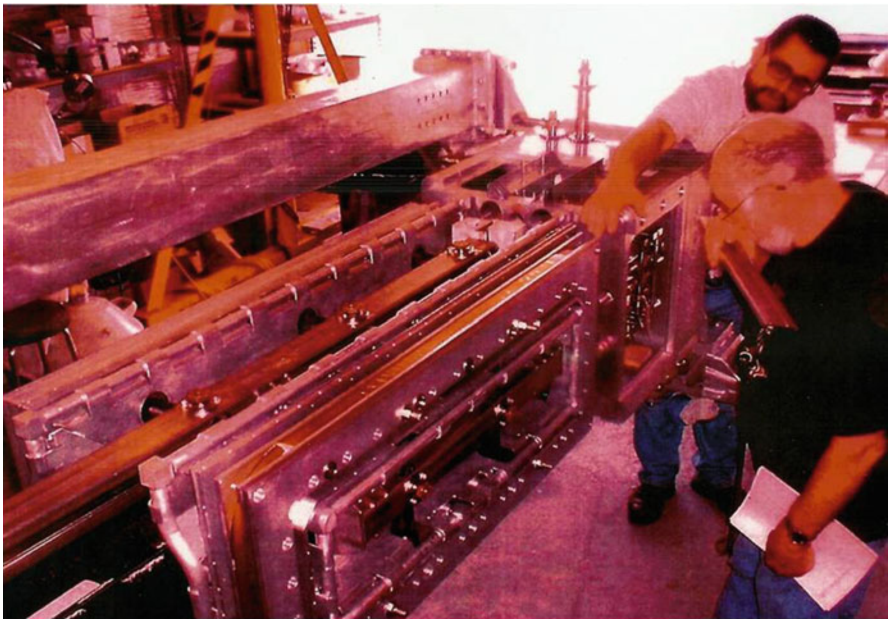
Figure 47 shows the completed SC quadrupole enclosed in its vacuum cryostat, while Fig. 48 shows testing of the quadrupole magnetic levitation and propulsion forces using DC current in the aluminum loop guideway assembly beneath the quadrupole as a stand-in for the induced currents. The quadrupole was successfully tested to its full design current of 600,000 A turns. The magnetic forces between the quadrupole and the guideway loop assembly were measured as a function of vertical separation and lateral displacement from the centered position, and



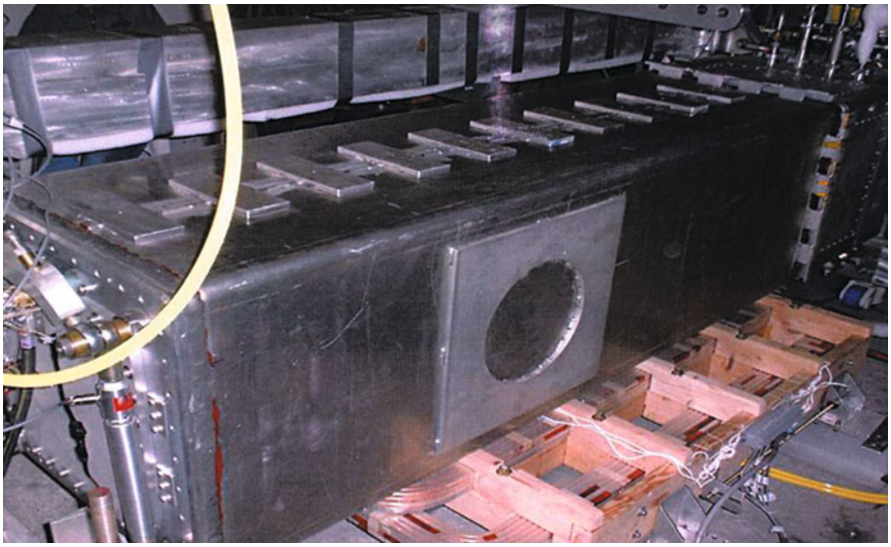
MAGLEV Technology Development. Figure 44
NbTi superconducting loop enclosed in stainless steel jacket



MAGLEV Technology Development. Figure 45
CAD-CAM drawing of Maglev-2000 superconducting quadrupole



MAGLEV Technology Development. Figure 46
Assembly of Maglev-2000 superconducting quadrupole



MAGLEV Technology Development. Figure 47
Completed Maglev-2000 quadrupole enclosed in its cryostat

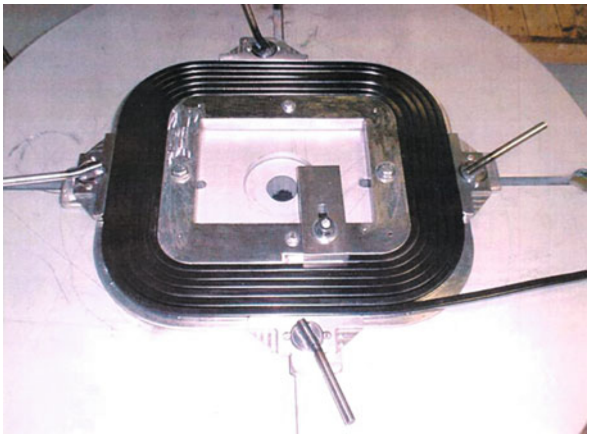
longitudinal position in the direction of movement along the guideway. The measured forces agreed with three-dimensional computer analyses.

In the time following the Maglev-2000 quadrupole tests, high temperature superconductors have become much more capable, and are commercially produced in substantial amounts. Using YBCO high temperature superconductor wire, it appears very possible to fabricate Maglev-2000 quadrupoles that would be much simpler in construction, with much easier refrigeration

MAGLEV Technology Development. Figure 48
Testing of magnetic forces on Maglev-2000 quadrupole using DC current in aluminum loop panel

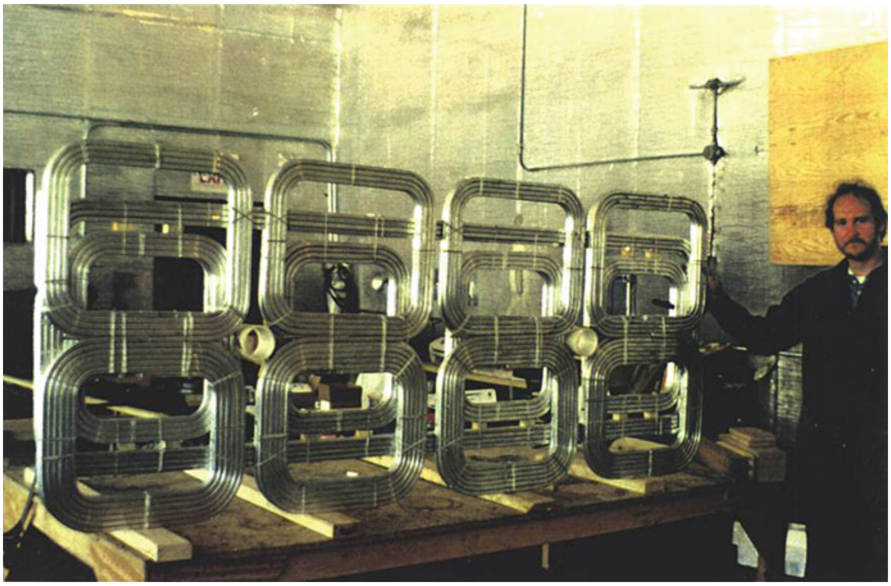
requirements. The YBCO superconductor would operate at 40 K with a much simpler on-board cryocooler than for NbTi superconductor at 4.2 K. One Maglev-2000 quadrupole requires 3,600 kA turns of superconductor. At $10 per kiloamp meter, which appears achievable with large-scale production of high temperature superconductor, the superconductor for it would cost $36,000. A passenger vehicle with eight quadrupoles would then have a superconductor cost of $288,000, while a truck carrying a vehicle with 16 quadrupoles would then have a superconductor cost of $576,000, both of which are very reasonable for a projected total cost of $5 million per vehicle. Future tests of Maglev-2000 quadrupoles will probably involve high temperature superconductors rather than NbTi superconductor with liquid helium coolant.

As described previously, the guideway loop panels contain three sets of wound aluminum loops, composed of a set of four figure of 8 loops, a set of four dipole loops, and one long LSM propulsion loop. Figure 49 shows a wound dipole loop, to be used in the panel. The aluminum conductor has a ∼10 mil layer of nylon using a dip process to coat the conductor. The nylon insulation withstood 10 kV tests without breakdown. Figure 50 shows a completed guideway loop panel with all of its loops.

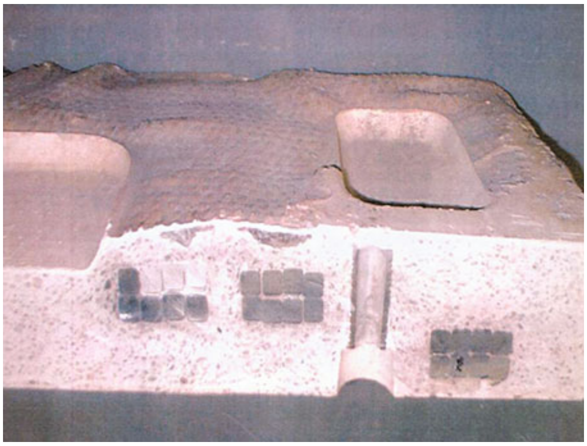


MAGLEV Technology Development. Figure 49
Wound dipole loop for guideway panel using nylon-coated aluminum conductor

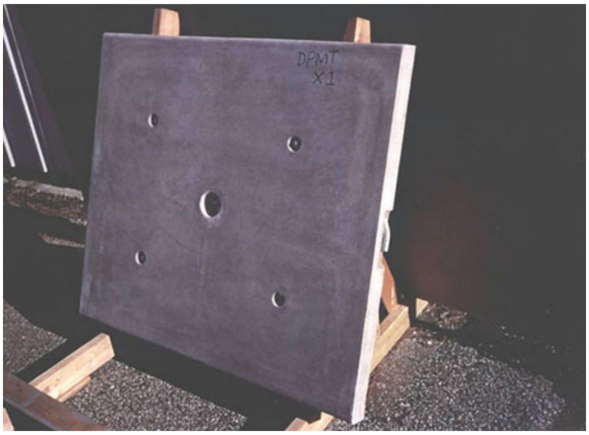
The completed panel is then enclosed in a polymer-concrete structure for handling and weather protection (Fig. 51). Polymer concrete – a mixture of aggregate, cement, and plastic monomer – can be cast into virtually any form as a slurry. When the monomer polymerizes (the rate of polymerization is controlled by the amount of added promoter), the resulting concrete-like structure



MAGLEV Technology Development. Figure 50
Completed guideway panel with figure of 8 dipole, and LSM propulsion loops



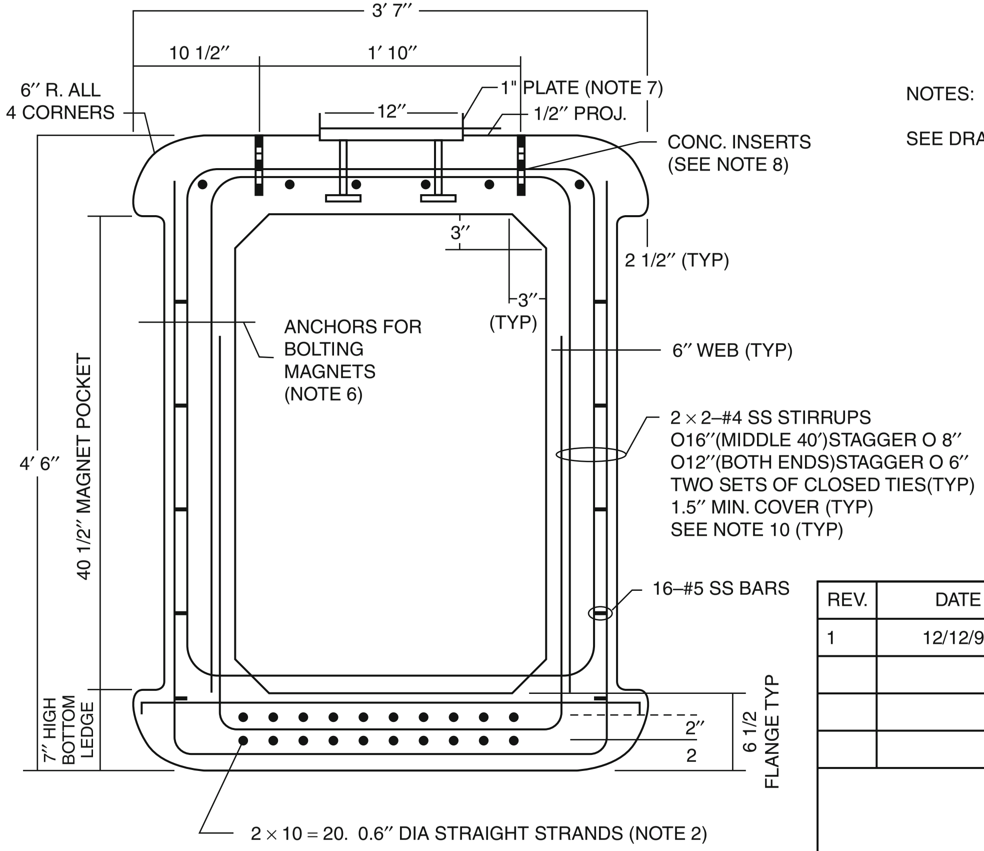
MAGLEV Technology Development. Figure 51
Guideway loop panel enclosed in polymer-concrete matrix



MAGLEV Technology Development. Figure 52
Polymer-concrete panel with enclosed aluminum loop exposed for 2 years to outdoor environment with multiple freeze-thaw cycles

is much stronger – a factor of 4 or greater – than ordinary concrete, and not affected by freeze-thaw cycles, salt, etc. Figure 52 shows a completed polymer-concrete panel left outside of the Long Island facility for 2 years. It was subjected to a wide range of weather conditions and multiple freeze-thaw cycles over the 2 year period, without any degradation.

After being fabricated at the Maglev factory, the guideway panels would be attached to the sides of the monorail or the surface of planar guideway beams to be shipped to a construction site for an elevated guideway, or transported to existing RR trackage that was to be modified for use by Maglev-2000 vehicles.



NOTES:

SEE DRAWING NUMBER 2000–2 FOR NOTES.

MAGLEV GUIDEWAY CROSS–SECTION
SCALE: 1″ = 1″– 0″

APPROX. ESTIMATED QUANTITIES:
CONCRETE: 21.12 CY; SS REBARS: 2,310 lb.
STRANDS: 1070 lb; WT OF BEAM: 86.72 KIPS

REV.	DATE	DESCRIPTION
1	12/12/98	CLOSED STIRRUP LOOPS

MAGLEV-2000 NARROW BEAM
72″ – 2 1/8″ LONG

DESIGNED BY	M. REICH, Ph. D.	7/13/98
CHECKED BY	A. CHIDAMBARAM, P. E.	7/13/98
DRAWN BY	W. GROSSMAN, P. E.	7/13/98
SHEET 1 OF 6	DRAWING NUMBER 2000–1	

MAGLEV Technology Development. Figure 53
Design for 22 m long monorail guideway beam

Figure 53 shows the basic design for the monorail guideway beam. It is a hollow box beam made with reinforced concrete. The beam length is 22 m and weight is 34,000 kg. It uses post-tension construction, which allows the tensioning cables in the base of the beam to be retightened if some stretching were to occur. The beam is tensioned to have a 0.5 cm upward camber at the midpoint of the beam when it is not carrying a Maglev vehicle. When the Maglev vehicle is on the beam, the beam flattens out to a straight line condition, with no vertical dip or camber along its length.

Figure 54 shows a photo of the fabricated beam after transport by highway truck from the manufacturing site in New Jersey to Maglev-2000's facility in Florida. No problems in the 800 mile transport by highway were encountered. The first beam cost $45,000 with a projected large-scale production cost of $25,000 per beam. Since 1999, construction costs have increased. At $50,000 in today's dollars, 140 beams for a two-way monorail guideway would cost $7 million per mile.

Figure 55 shows a CAD-CAM drawing of the aluminum chassis that was fabricated for a 20 m long Maglev-2000 test vehicle, designed to carry 60 passengers in urban and suburban service. Figure 56 shows the fuselage for the test vehicle.

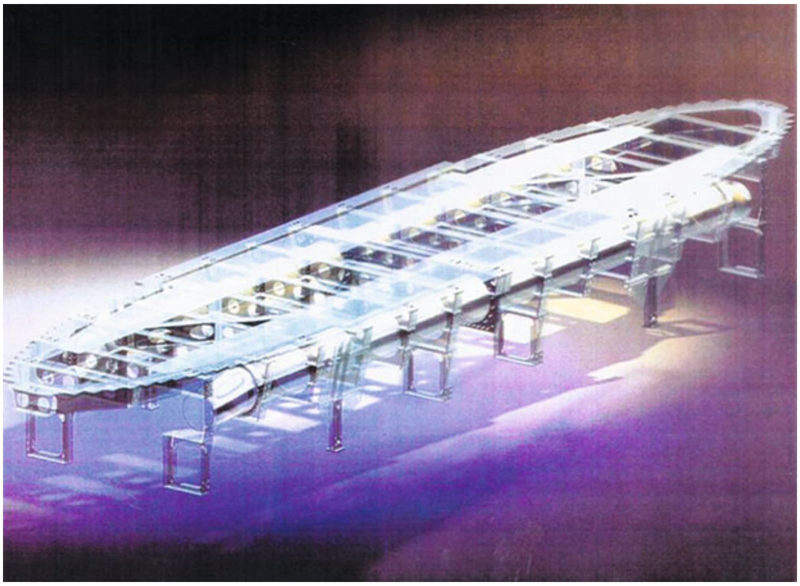
Maglev Applications

High Speed National Maglev Networks

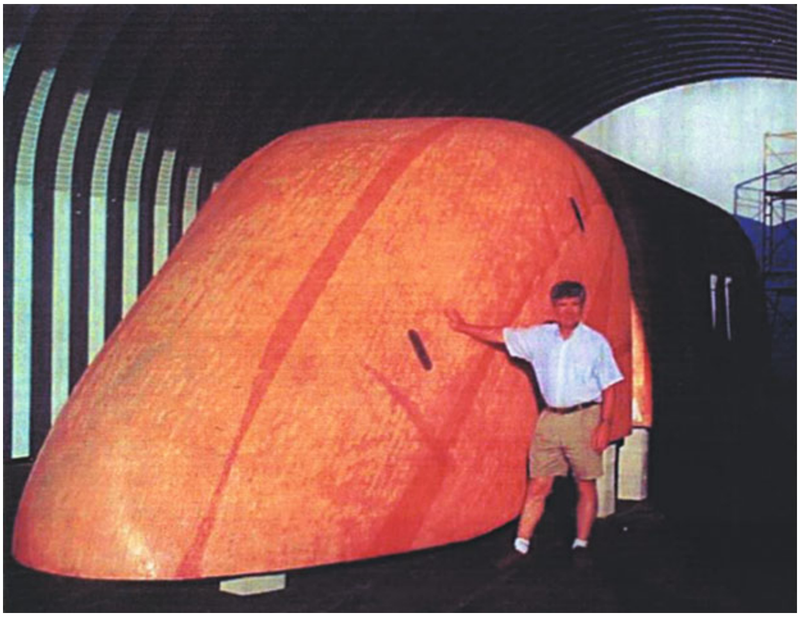
As in other developed nations, USA's three main transport systems – motor vehicles (autos, buses, and trucks), airplanes, and conventional rail – operate as national networks. From any given area in the USA, passengers and freight can drive, fly, or go by bus or train to any other area in the USA. To reach a particular location, it may be necessary to transition from one system to another; for example, one can fly from one airport to another airport, with a short drive to one's final destination. However, such local transitions are usually easily accommodated.

For Maglev to be an important mode of transport in the twenty-first century, it must also function as a national network. A few isolated Maglev routes, while helpful to local travelers, will provide only minor benefits. They will not substantially reduce oil consumption and greenhouse gas emissions, and not significantly increase domestic jobs and exports.

MAGLEV Technology Development. Figure 54
Photo of 22 m long monorail guideway beam delivered to Maglev-2000 facility in Florida from construction site in New Jersey

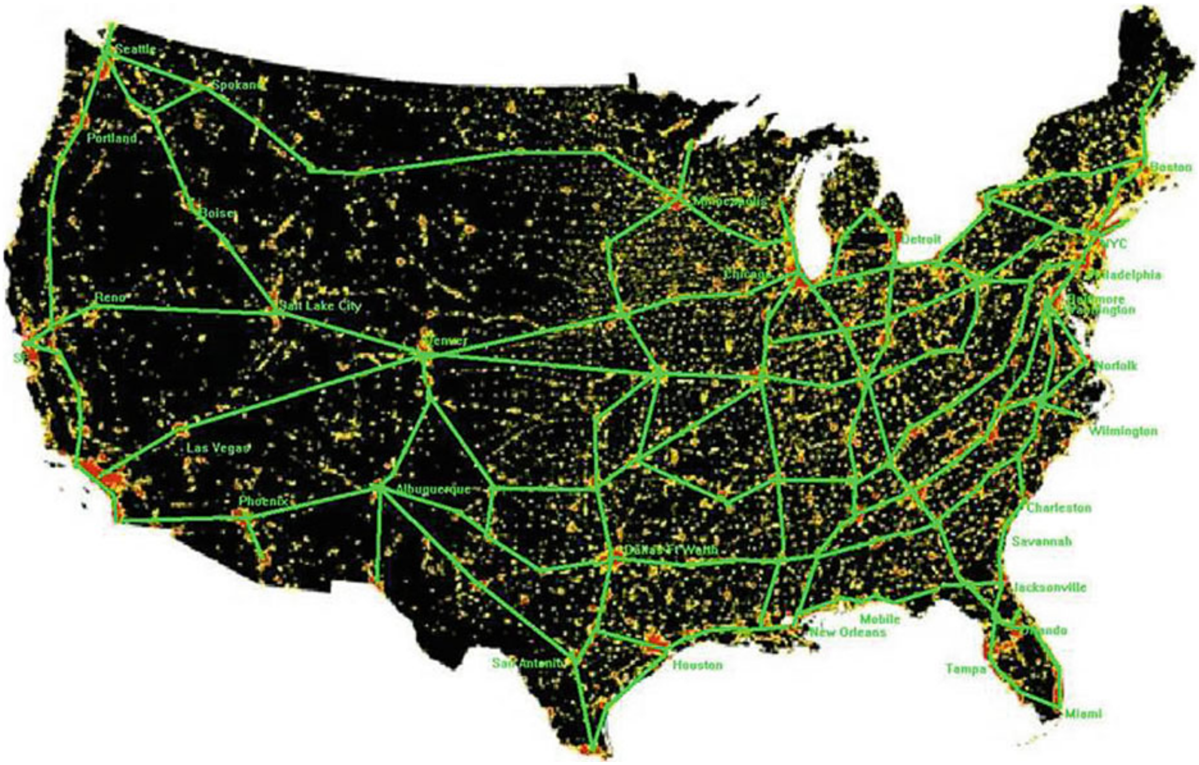


MAGLEV Technology Development. Figure 55
CAD-CAM drawing of aluminum chassis for 60 ft long Maglev-2000 vehicle



MAGLEV Technology Development. Figure 56
Photo of fuselage for 60 ft long Maglev-2000 vehicle

The potential 25,000 mile National Maglev Network (Fig. 57) would interconnect virtually all major metropolitan areas in the USA. The intercity Maglev routes, following the vision of the late Senator Daniel Patrick Moynihan, would primarily be on the rights of way of the Interstate Highway System [26]. Had his $750 million 1990 Senate passed legislation for a US Maglev R&D program not been killed in the House of Representatives, USA would be well on its way to its National Maglev Network.



MAGLEV Technology Development. Figure 57
25,000 mile national Maglev-2000 network

One of the unique features of the second-generation Maglev-2000 system is the ability of Maglev-2000 vehicles for levitated travel along existing RR tracks, to which thin panels holding aluminum loops have been attached to the cross-ties.

Conventional trains can continue to use the RR tracks, given appropriate scheduling. This capability to use existing RR tracks enables Maglev-2000 vehicles to travel in densely populated urban and suburban areas, without needing to construct very expensive new infrastructure, with its inevitable disruptions to the existing infrastructure and the local population. Combined with high speed elevated Maglev guideways between metropolitan areas, this would result in fast convenient travel by Maglev, both within a given metropolitan area, and from one metropolitan area to another.

In addition to the attractive transport features and economic benefits of the National Maglev Network benefits and compared to present transport systems, its ability to effectively and cheaply transport passengers and freight without the need for oil will become extremely important in the following decades as world oil runs out.

The desired capabilities for a US National Maglev Network include:

1. Low guideway cost
2. Able to transport passengers, personal autos, highway trucks, and freight containers on dual-use guideways with high energy efficiency in all weather conditions
3. Able to be privately financed without the need for government funding and subsidization for construction and operation
4. Rapid installation of guideways with minimal disruptions and modifications to existing infrastructure
5. High speed electronic switching to off-line stations, with service to multiple convenient stations in metropolitan areas
6. Earth ambient magnetic field levels in passenger cabins

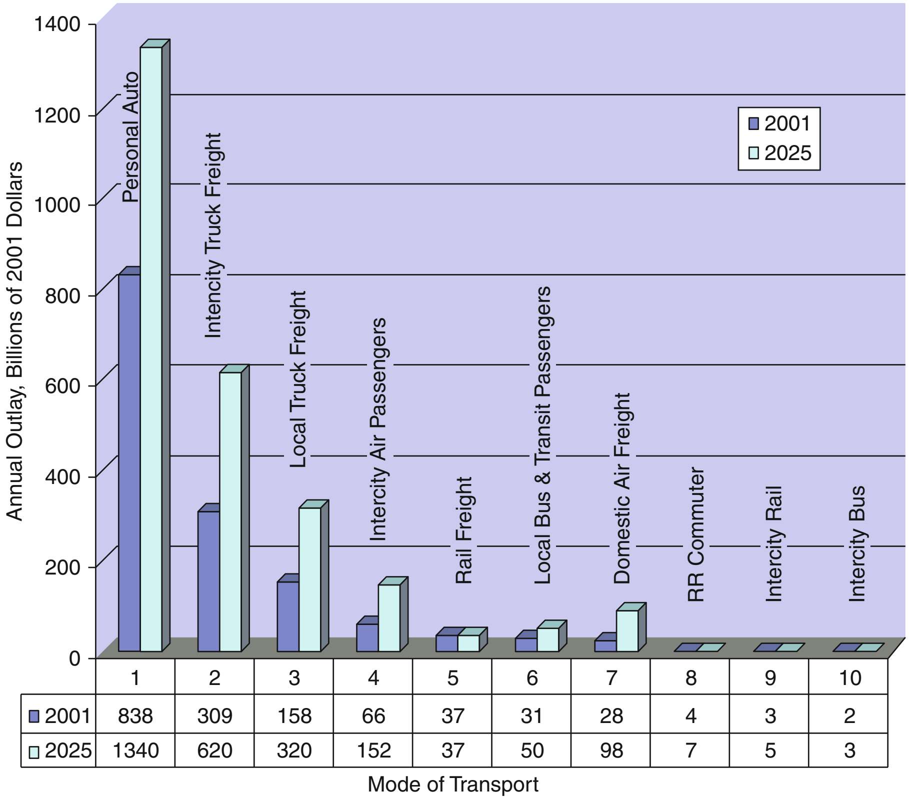
Construction costs for the Japanese and German first-generation Maglev systems are high, for example, $50 million or more per two-way mile [27]. To be widely implemented in the USA, construction cost should be substantially lower (Capability #1).

The National Maglev Network would not just carry passengers, but also intercity highway trucks, personal autos, and freight containers (Capability #2). US transport outlays [24] for intercity highway trucks (over $300 billion) are much greater than those for passenger air ($60 billion annually) and dwarf passenger rail (only $3 billion per year) (Fig. 58).

Intercity highway trucks haul about the same amount of freight, 1.5 trillion ton miles per year as conventional rail, but charge ten times as much per ton mile transported. They are preferred for high-value freight transport over conventional rail because of their much faster delivery and their convenient service, direct from origin to destination. Shipping by rail takes much longer and is much less convenient. The average haul distance for highway trucks is ∼500 miles. Truckers would need far fewer trucks to deliver the same volume of goods if they went by Maglev at 300 mph rather than by highway at 60 mph, and at less cost per ton mile delivered (Fig. 59).

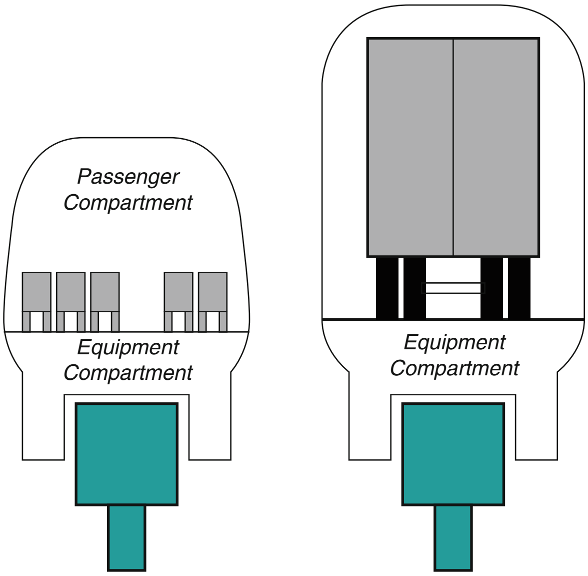
Similarly, drivers could take their personal autos with them on Maglev vehicles configured to carry 15–20 autos and their passengers. Travel would be much faster and cheaper than by highway, counting fuel, lodging, and auto maintenance costs.

Private financing is essential for a National Maglev Network (Capability #3) because of the very high government debt levels. Passenger-only Maglev systems will not attract private financing. Figure 60 shows the payback time for a Maglev route as a function of the number of trucks transported daily, for the operating parameters summarized in Table 2 assuming zero passenger revenues (curve A). At 3,000 trucks daily (20% of typical highway truck traffic on an interstate) the payback time is less than 5 years. At 6,000 trucks daily



	1	2	3	4	5	6	7	8	9	10
■ 2001	838	309	158	66	37	31	28	4	3	2
□ 2025	1340	620	320	152	37	50	98	7	5	3

MAGLEV Technology Development. Figure 58
Annual outlays current and future, for US transport modes (billions of dollars) (US census Bureau [24])

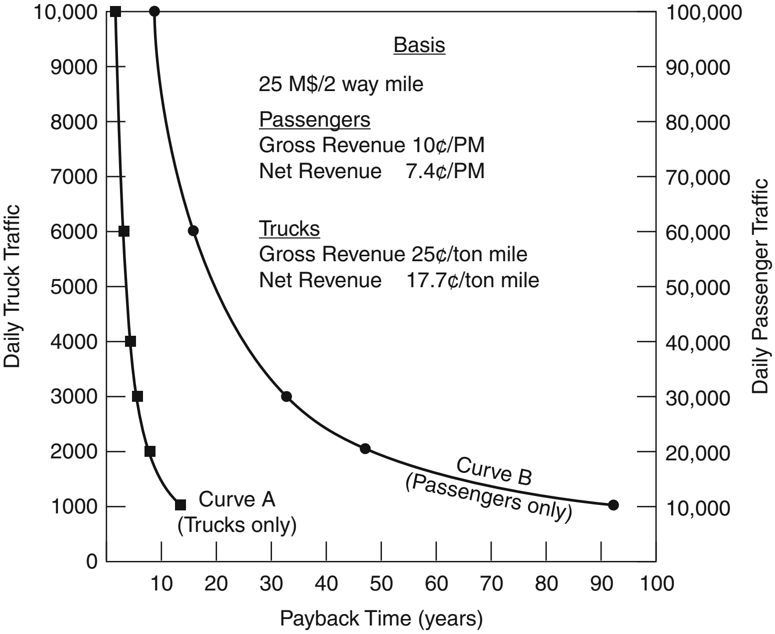


MAGLEV Technology Development. Figure 59
Maglev-2000 system can handle both freight and passengers

(40% of truck traffic) payback time is less than 2.5 years. Without trucks, payback time is many decades for realistic passenger ridership levels of 10,000–20,000 passengers daily (curve B).

Rapidly installing guideways with minimal disruption and modification of existing infrastructure (Capability #4) minimizes the cost and local opposition to the construction of a Maglev route. This is especially important in densely populated urban and suburban areas. Whenever possible, existing infrastructure including existing rail trackage should be used to avoid disruptions to local inhabitants.

High speed electronic switching (Capability #5) is very desirable. Not being able to switch off the main line, or having to mechanically move a long cumbersome guideway switch section significantly reduces average speed. For example, HSR systems, though capable of maximum speeds of ~200 mph, have considerably slower average speeds when station stops and their slow acceleration and deceleration rates are taken into account [28]. Electronic switching to off-line stations will enable Maglev vehicles to bypass at high speed stations they are not scheduled to stop at, maintaining high average speeds. This will allow many



MAGLEV Technology Development. Figure 60
Payback time for Maglev-2000 guideway as a function of truck and passenger traffic

MAGLEV Technology Development. Table 2 Vehicle O&M costs

Revenues and costs	Passengers (cents/pm)	Trucks (cents/ton mile)
Gross rev	10	25
Energy cost	1.2	4.0
Am&M cost	0.9	2.8
Personnel cost	0.5	0.5
Net Rev.	7.4	17.7

$5 million vehicle cost; 10 year amortization; 5%/year maintenance
100 passenger or 30 t capacity; 80% load factor; 12 h op/day
250 mph average speed; 3 MW propulsion power for passenger vehicles
4 MW for trucks; 6 cents/KWH

MAGLEV Technology Development. Table 3 National network and Maglev-2000 system parameters

Network	25,000 miles, 300 mph max
Passenger vehicles	100 passengers
Truck carriers	Two types (one and two trucks)
Auto carriers	15 autos + passengers
Travel	Either as single units or multiunit consists, depending on traffic
Quadrupole magnets	600,000 A turns 18 in. wide, hi temp superconductor, 8 magnets for passenger vehicle, 16 magnets on truck and auto carriers
Magnetic suspension	10 cm gap between vehicle and guideway, 0.3 g/cm automatic magnetic restoring force

closely spaced stations within a metropolitan area, for convenient access.

Earth ambient magnetic field levels in passenger compartments (Capability #6) are not only desirable, but a necessity. The US public strongly resists new technologies that appear to deviate from their normal environment.

There is no evidence that DC magnetic fields of a few gauss have any effect on the body (there is a limit of 5 gauss for people with pacemakers) and people experience much greater fields – at the kilogauss level during MRIs – without problems. However, to avoid possible controversy and opposition to Maglev, it is desirable to keep the magnetic field inside passenger cabins at Earth ambient level.

The guideway panels can also be mounted on the cross-ties of existing RR tracks (Fig. 40), enabling levitated travel of Maglev-2000 vehicles along existing RR tracks in a planar guideway mode. The panels do not interfere with the operation of conventional trains, which could continue to use the tracks for bulk freight transport, given appropriate scheduling, probably at nighttime. The ability of Maglev-2000 vehicles to travel as individual units would enable much more frequent and convenient passenger service, rather as long trains of many RR cars. Also, because the Maglev vehicle loads are distributed along the vehicle and not concentrated at wheels, its local track loading is much less than conventional trains, resulting in much longer track life and reduced maintenance.

The capital cost of the monorail guideway has been examined in detail, based on fabrication experience and costs of magnets, panels, and the guideway beam. In 2000 dollars, the projected cost was $11.4 million per two-way mile for "greenfields" construction (no land acquisition or infrastructure modification costs) [29]. In 2010 dollars, the corresponding construction cost would be ~$25 million per two-way mile. The cost to adapt existing RR tracks for Maglev-2000 operation would be much less, on the order of $4 million per two-way mile, since an elevated guideway would not be required.

Table 3 summaries the nominal operating parameters for the Maglev-2000 system. A key element of the Maglev-2000 system that results in low construction cost is to mass produce its prefabricated components in large factories. The components, beams with attached panels, piers, controls, etc., can then all be shipped to the construction site and quickly erected at low cost on pre-poured concrete footings using conventional cranes. The components can also be readily exported to other countries. A container ship, for example, can carry 20 miles of prefabricated two-way guideway.

The proposed 2000 mile West Coast Maglev Network (Fig. 57) along the I-5 corridor and its branches would connect the metropolitan areas in California, Nevada, Oregon, and Washington State, plus the Vancouver, British Columbia, area into a high speed Maglev system that would quietly transport many thousands of passengers, highway trucks, and personal autos 6 in. off the rails at speeds of 300 mph and one-fifth the cost of airline travel. Forty-two million persons in California, Nevada, Oregon, and Washington, 85% of the total population of 49 million in the four states, would live within 15 miles of their local Maglev station, from which they could reach any other station in the Maglev Network within a few hours. Another 1.3 million persons in the Vancouver area would be served by the network, making the total population within 15 miles of a Maglev station equal to 43.3 million people. Total government funding is limited to $600 million over 5 years for up front demonstration and certification activities. After that, freight capability enables building the entire network with private financing.

Typical trip times on the Maglev Network, compared to going by highway would be:

San Diego to Seattle	4 h 30 min vs 25 h 15 min
San Francisco to Los Angeles	1 h 45 min vs 9 h 40 min
Portland to San Francisco	2 h 30 min vs 12 h 45 min
Los Angeles to Las Vegas	1 h vs 5 h 30 min

In addition to much shorter trip times by Maglev, the cost of travel by Maglev would be significantly less for passengers, highway trucks, and personal autos as compared to existing transport modes:

Passengers – 3 cents per passenger mile (PM) on Maglev, compared to 40 cents per PM for driving by auto, 15 cents per PM by air, and 50 cents per PM by high speed rail.

Highway trucks – 10 cents per ton mile by Maglev compared to 30 cents per ton mile by highway.

Personal autos – 30 cents per mile by Maglev transport compared to 40 cents per mile by highway.

Table 4 gives data on the vehicle flow and congestion along the I-5 corridor.

Based on the data provided in Table 4 [30], a reference case for the average flow on the West Coast Maglev Network was assumed, with

- 5,000 trucks per day (2 per Maglev vehicle)
- 30,000 passengers per day (100 per Maglev vehicle)
- 20,000 personal autos carried on Maglev per day (10 per Maglev vehicle)

MAGLEV Technology Development. Table 4 Vehicle flow and congestion along the I-5 corridor

Traffic flow on corridor				
	2007		2035	
	Avg.	Max	Avg.	Max
Vehicles/day	71,000	300,000	150,000	~600,000
Trucks/day	10,000	35,000	22,000	~70,000
Urban segments,[a] % congestion	65%		95%	
Rural segments, % congestion	31%		85%	

[a] 550 miles of 1,381 mile total length are urban segments

The above traffic flows on Maglev are a significant fraction of the 2035 traffic flow, but could be increased substantially. For the reference case, travels by Maglev results in the major benefits given below:

- $21 billion annual reduction in cost of transport
- 53 billion KWH annual reduction in energy for transport
- 27 million tons annual reduction in CO_2 emissions

The West Coast Maglev Network can be in full operation by 2019, given an aggressive program with adequate funding. Construction ground breaking would start in 2015, following a 5 year program to demonstrate and certify the various types of Maglev vehicles – passenger, auto with passengers carrier, and highway truck carrier on a test guideway at commercial operating conditions. Planning activities, environmental permits, and startup of guideway production plants would also be carried out in parallel with the testing of commercial-type vehicles. The network would be constructed over a 4 year period by simultaneously carrying out construction of nine segments of the 2,000 mile network. Each segment would run between major metropolitan areas on the network, for example, San Diego to Las Angeles would be one segment, Anaheim to Las Vegas another segment, and so on. Each segment of the Network would have one or more guideway beam and pier manufacturing plants. The projected construction cost is on the order of $60 billion, based on an average unit cost of $25 million per two-way mile.

Maglev Energy Storage

Electrical power systems typically experience large swings in power demand, with low demand during the night and high demand during the daytime. Even in the daytime, power demand varies widely, with major peaks during the morning and late afternoon. Existing technologies for electrical power storage are generally too expensive and difficult, for example, batteries, flywheels, superconducting energy storage (SMES), or too limited in siting, for example, pumped hydro. Today, the great majority of peak power demand is supplied by fossil fueled peaking power plants, for example, gas turbine, or by purchase from distant power grids.

A new approach for the storage and rapid delivery of electric power, based on Maglev, is proposed [31]. This new method, termed *MA*glev *P*ower *S*torage (MAPS) uses Maglev technology similar to that employed for the transport of passengers and freight to lift masses a kilometer or more in altitude. In doing so, electrical energy is stored as gravitational potential energy. Raising a 100 t mass 2 km in altitude, for example, stores 0.5 Megawatt Hour (MWH) of energy. Raising 2,000 such masses would store 1,000 MWH of energy, which would be returned to the power grid at appropriate times by simply transporting the masses down to a lower altitude. During periods of electrical storage, Maglev vehicles would transport masses up to a higher altitude storage facility, with the vehicles' magnetic propulsion system operating in the motor mode. During periods of electrical power delivery, the masses would be transported down to lower altitude, with the propulsion system operating in the generator mode.

The total energy delivered is determined by the number of storage masses transported up and down, while the power delivered is determined by the rate at which they are transported. The MAPS system is very flexible and can rapidly alter its power level by changing the rate at which masses are moved up or down the guideway. MAPS power level, for example, could go from zero to 100% of full capability, or from 100% to zero, in less than 2 min. Peak power capabilities of ~1,000 MW(e) can be generated using MAPS. Because of its rapid response capability, in addition to providing peak power, MAPS system could also be used for low cost spinning reserve. Presently, spinning reserve is provided by keeping expensive generator units hot and ready to generate large blocks of power .

While the basic Maglev technology for energy storage is similar to that for passenger and freight transport, there are differences, that is, steeper grades for the MAPS system (up to 45° vs a maximum of ~10° for passenger/freight transport), and considerably heavier vehicles (~100 t vs a maximum of ~50 t for freight transport). The heavier loaded and much shorter vehicles (~6 m for MAPS vs ~30 m for passenger/freight vehicles) use a magnet configuration which maximizes propulsion force relative to I^2R losses in the propulsive windings. To efficiently levitate the short heavy vehicle with minimum I^2R losses, iron plates in the narrow beam guideway provide most of the levitation force, with null flux loops for stability and oscillation damping. The storage masses are simple reinforced large concrete blocks on the flat upper surface of the MAPS vehicle. Typical dimensions for a 100 t storage mass are 3.5 m wide, 2.5 m high, and 4.5 m long. The blocks can be rapidly moved onto or off from the vehicle using a roller bar surface and stored on pads at the guideways high and low altitude points. A 100 m × 600 m long (6 acres) storage facility could handle and deliver 2,000 storage masses, equal to 1,000 MWH of electrical energy at an elevation change of 2 km (6,000 ft). Overall energy efficiency, output electrical energy/input electrical energy, would be well above 90% – much higher than any other electrical storage technology. For a mass unload time of 50 s and a mass load time also of 50 s (both steps are necessary during a vehicle round trip, whether MAPS is operating in the power storage or power delivery mode) a given MAPS vehicle could make 20 round trips per hour. This is quite conservative, since the 50 s intervals could probably be reduced to about 30 s, considering the simplicity of the transfer process. The total number of round trips per hour would then increase.

A very attractive feature is the completely nonpolluting environmentally benign nature of MAPS, and its ability to store power at very low cost from renewable wind and solar energy sources. MAPS facilities could be built in a short time, for example, 3 years or less, to meet increasing peak power demands.

The total land use for a 2,000 MWH facility, including a dual 3 km long guideway, is only about 20 acres. Most MAPS facilities would be sited in remote area and would not impact the environment. At the design speed of 130 mph, the vehicles would be virtually silent.

In locations with flat terrain, MAPS guideways could be installed underground in inclined tunnels or vertical shafts. Many mining tunnels and shafts operate at depths of 1 km (3,300 ft) or more. MAPS would transport its concrete blocks between a storage yard on the surface, and underground storage cavities positioned along a guideway in an underground horizontal tunnel. There are many mines, both coal and hard rock, in the USA that are no longer in use. Such mines would be attractive sites for MAPS facilities in flat regions of the USA.

Maglev Water Transport

The availability of ample clean water is a major concern in many regions, and is expected to be even worse in the decades ahead – in fact, many experts believe it to be the number one problem facing the world. A large fraction of the world's population, most of them poor, already lives in a state of water scarcity. World population is expected to grow to nine billion (mid-range projection) by 2050. This growth, along with increasing industrialization, urbanization, and irrigation, will put even more stress on water resources.

The advanced second-generation Maglev-2000 system described previously is designed to transport passengers and freight at very low cost. With design modifications to the vehicles and guideway, the second-generation Maglev-2000, termed the Water Train [32] system, can transport large amounts of water, that is, on the order of one billion gallons daily (1,000 megagallons/day), for hundreds of miles.

The Water Train vehicles transport a higher load per unit length of vehicle than for passenger and truck transport. The weight of loaded Maglev passenger/freight vehicles is nominally 50 t for a 30 m length. A water transport vehicle will probably weigh 200–300 t for the same length. This much heavier loading requires that the guideway be on-grade, or if it utilizes discrete supports, that they be relatively close together, for example, every 20 ft or so. The Water Train uses the "iron lift" configuration proposed by Danby and Powell [33], in which superconducting magnets on the vehicle are attracted upward to laminated steel plates attached to the guideway. The suspension is designed so there is a vertical equilibrium point at which the vertical lift force from the steel plates equals the weight of the vehicle.

The "iron lift" configuration is vertically stable, since as the vehicle moves upward, the lift force becomes smaller, and as it moves downward, the lift force becomes larger. The iron plates, however, make the "iron lift" configuration horizontally unstable. This horizontal instability is countered by the null flux stability loops on the guideway. The stable restoring force from the null flux loops is greater than the unstable force from the attraction to the steel plates, making the suspension horizontally stable.

An important feature of the Water Train is the flexible "balloon" that carries the water on each vehicle. The "skin" of the balloon is a relatively thick, for example, 1 in., reinforced flexible composite material. When pressurized and filled with water, the container forms a streamlined cylinder that runs the full length of the vehicle. After delivery of the water load and depressurization, the flexible skin is depressurized and collapsed minimizing the frontal area and air drag on the vehicle. The air drag resistance for a long multi-vehicle train will be much smaller than for a single vehicle. The equivalent drag per vehicle of the multiunit train is only 24% of a single vehicle.

The three principal cost components for the Water Train are propulsion energy and amortization of the vehicles and guideway. Energy cost scales as V^2, and vehicle amortization scales as 1/V. The faster the vehicle travels, the more water it can carry in a given period of time, reducing its amortization cost. Guideway cost is independent of vehicle speed, and depends only on the total amount of water carried. For a 1,000 megagallon per day delivery rate – the rate for a large city – amortization of the 300 mile long guideway only costs about 16 cents per 1,000 gal. Adding vehicle amortization and propulsion energy costs, the total cost for a 300 mile delivery is only about $1.50/1,000 gal. If the Water Train delivers water from a higher altitude to a lower one, part or all of the energy cost would be offset by gravitational energy input from the decrease in altitude. A 600 m drop from the origin to its destination would enable the transport of water at zero energy cost. Over the 500 km delivery distance, this is an average grade of about 1 m/km – less than many natural river systems. A greater drop in altitude would enable the sale of surplus electric power generated by the moving vehicles to the electric grid. Accordingly, the Water Train can also generate large amounts of hydroelectric power at sites where there is substantial water flow and rapid drop in elevation. As an alternative to large dams, such an electric generation system would have significant advantages, including a much lower capital cost and much smaller environmental impact.

Elevation changes in terrain traversed by the Water Train are easily accommodated by small changes in train speed. Even for changes of ±200 m (660 ft), the velocity changes are modest and acceptable, without the need for propulsive power. The Water Train would simply speed up slightly as it coasted downhill, or slow down slightly as it coasted uphill, depending on the local terrain.

Finally, in contrast to long distance pipelines, which require many pumping stations along their length, the Water Train, because of its high kinetic energy and low air drag, can coast for hundreds of miles without additional propulsion. After the Water Train attains an initial velocity of 100 m/s (225 mph), and coasts for a distance of 100 miles, the Water Train velocity has dropped to 80% of the initial value, with the average velocity over the 100 mile segment at ~90% of its initial speed. While the Water Train could coast the entire 300 miles without propulsive input, it probably would be desirable to reaccelerate at intervals of 100–150 miles in order to keep the average velocity close to the optimum for minimizing cost. At its destination, the Water Train would electromagnetically de-accelerate down to zero velocity, and its kinetic energy fed back as electric power into the grid.

Future Directions

Maglev is still at an early stage of development, both in its technology and in its various applications. In the technology area, Maglev will evolve to use high temperature superconductors (HTS) that will enable operating temperatures in the range of 20–40 K, with much lower refrigeration power requirements, that is, W(e) of refrigeration electric power per watt(th) of thermal input through the cryostat insulation, than those for the present liquid helium cooled superconductors that operate at 4.2 K. HTS superconducting Maglev magnets will require only one-tenth as much refrigeration power input as liquid helium cooled magnets.

The lower refrigeration power and greater simplicity of HTS magnets will enable Maglev vehicles to be smaller, carry fewer passengers, and serve a wider range of transport applications. Maglev vehicles can be adapted for levitated travel in existing subways and commuter rail lines, providing faster and more frequent service in urban and suburban areas.

Maglev platforms will be developed to provide faster, easier, and safer movement of heavy objects in industrial plants, and movement of ores dug from open pits and underground mines at lower cost. Maglev vehicles will provide fast, convenient, and nonpolluting movement of heavy freight containers to and from seaports, in place of the thousands of greenhouse gas–emitting trucks that now provide the service.

Bibliography

1. O'Toole R (2009) "Gridlock: why we are stuck in traffic and what to do about it." CATO Institute
2. http://www.eia.doe.gov/bookshelf/brochures/greenhouse/Chapter1.htm
3. http://en.wikipedia.org/wiki/List_of_Countries_by_canbon dioxide_emissions
4. http://en.wikipedia.org/wiki/chevrolet_volt
5. http://en.wikipedia.org/wiki/Transrapid
6. (1914) The Bachelet levitated railway. Nature 93(2324):273
7. Danby G, Powell J (2000) Magnetic levitation: a new mode of transport for the 21st century, lecture given at the award for the Franklin Medal for Engineering, Franklin Institute, 26 Apr 2000
8. Powell J, Danby G (1966) High speed transport by magnetically suspended trains. Paper 66-WA/RR-5, presented at 1966 Winter ASME Meeting, New York City
9. www.rtri.or.JP/rd/maglev/html/english/maglev_frame.E.html
10. http://en.wikipedia.org/wiki/vactrain
11. http://magnetbahnforum.de/index.php?_mf_steering_committee
12. http://en.wikipedia.org/wiki/Heike_Kamerlingh_Onnes
13. http://en.wikipedia.org/wiki/Large_Hadron_Collider
14. Census Bureau (2006) US statistical abstracts, Table 1046, transportation outlays by type of transport: 1970 to 2001. Census Bureau, Washington, DC
15. Powell J, Danby G (1967) High speed transport by magnetically suspended trains. Paper 66-WA/RR-5, presented at 1966 Winter ASME Meeting, New York City; Mech Eng 89:30–35
16. Kolm H et al (1975) The magneplane system. Cryogenics 15:377–384

17. http://en.wikipedia.org/wiki/Maglev_(transport)
18. http://www.upf.org/bering-strait.../101.../479-Japanese-maglev-technology
19. http://en.wikipedia.org/wiki/Shanghai_Maglev_Train
20. News.smh.com.au ≥ Technology
21. http://www.rtri.or.jp/rd/maglev/html/english/maglev-frame_E.html
22. http://www.digitalworldtokjo.com.../500kph_maglev_flying_trains_get_green_light_to_burn_up_japan
23. http://magnetbahnforum.de/index.php?speed-records
24. US Census Bureau (2006) Statistical abstracts of the United States 2006, tables 1046 and 1047. US Department of Commerce, Washington, DC
25. Powell J et al (2008) Fabrication and testing of full scale components for the 2nd generation Maglev-2000 system. In: Maglev 2008 Conference, San Diego
26. Powell J, Danby G (eds) (1989) Benefits of magnetically levitated high speed transportation for the United States, report to United States Senate Committee on environment and public works, Grumman corporation. Bethpage, New York
27. Maglev I (2005) The Pennsylvania project: FRA high speed Maglev deployment program – project description. Transrapid International and Maglev, Pittsburgh
28. http://www.o-keating.com/hsr/tgv/htm
29. Powell J (ed) (1999) Cost projections for the Maglev-2000 system. Report DPMT-20, Maglev 2000 of Florida
30. http://www.thwa.dot.gov/pressroom/dot0795.htm
31. Powell J (2000) Electrical power storage and delivery using Maglev: the Maglev power storage system (MAPS). Report DPMT-14, DPMT Corp., 1 Nov 2000
32. Powell J, Danby G The water train: long distance delivery of water by Maglev. Report DPMT-3
33. Danby GT, Jackson JW, Powell JR (1974) Force calculations for hybrid (ferro-null flux) low drag systems. IEEE Trans Magnetic Mag O:443

Malaria Vaccines

Christopher V. Plowe
Howard Hughes Medical Institute/Center for Vaccine Development, University of Maryland School of Medicine, Baltimore, MD, USA

Article Outline

Glossary

Adjuvant A substance added to a vaccine to stimulate a stronger or more effective immune response.

Blood stage The stage of the malaria parasite life cycle responsible for clinical symptoms. Vaccines that target the blood stage are intended to prevent disease and death, but they do not prevent infection and may not affect malaria transmission.

Challenge trial Small experimental Phase 1/2 clinical trial in which healthy volunteers receive a malaria vaccine and are exposed to the bites of malaria-infected mosquitoes or injected with malaria parasites under carefully controlled conditions.

Immunogenicity The ability of a vaccine to produce specific immune responses (usually antibodies) that recognize the vaccine antigen.

Pre-erythrocytic Stages of the malaria parasite that are injected by a mosquito and develop in the liver before emerging into the blood where they can cause symptoms. Vaccines targeting pre-erythrocytic stages are intended to prevent infection altogether and, if highly effective, would also prevent disease and block transmission.

Sexual stage The male and female forms of malaria parasites that are responsible for transmission through mosquitoes. Vaccines directed against sexual stages are intended to prevent malaria transmission.

Subunit vaccine A vaccine based on a small portion of the organism, usually a peptide or protein.

Vaccine resistance The ability of malaria parasites to escape strain-specific immune responses by exploiting genetic diversity to increase the frequencies of non-vaccine-type variants in a population or to evolve new diverse forms.

Whole-organism vaccine A vaccine based on an attenuated or killed whole parasite.

Definition of the Subject

Vaccines are the most powerful public health tools mankind has created, and research toward malaria vaccines began not long after the parasite responsible for this global killer was discovered and its life cycle described more than 100 years ago. But parasites are bigger, more complicated, and wilier than the viruses and bacteria that have been conquered or controlled with vaccines, and a malaria vaccine has remained elusive. High levels of protective efficacy were achieved in crude early experiments in animals and humans using weakened whole parasites, but the results of more sophisticated modern approaches using molecular techniques have ranged from modest success to abject failure. A subunit recombinant protein vaccine that affords in the neighborhood of 25–50% protective efficacy against malaria is in the late stages of clinical evaluation in Africa. Incremental improvements on this successful vaccine are possible and worth pursuing, but the best hope for a malaria vaccine that would improve prospects for malaria eradication may lie with the use of attenuated whole parasites and powerful immune-boosting adjuvants.

Introduction

The malaria parasite is thought to have killed more human beings throughout history than any other single cause [1]. Today, along with AIDS and tuberculosis, malaria remains one of the "big three" infectious diseases, every year exacting a heavy toll on human life and health in parts of Central and South America, large regions of Asia, and throughout most of sub-Saharan Africa, where up to 90% of malaria deaths occur [2].

In the middle of the twentieth century, the availability of the long-acting insecticide dichlorodiphenyltrichloroethane (DDT) and the safe and effective antimalarial drug chloroquine provided the basis for optimism that global eradication was possible. Although malaria was eliminated in several countries around the margins of the malaria map, factors including the emergence of DDT-resistant mosquitoes and chloroquine-resistant parasites, as well as waning economic and political support, led the campaign to stall by the late 1960s, resulting in the rapid resurgence to previous disease levels in many locations [3].

For the next 30 years, the spread of drug-resistant malaria and donor fatigue contributed to an overall lack of progress against the disease. But starting in the late 1990s, new tools, including long-lasting insecticide-impregnated nets and highly efficacious combination drug therapies, led to a wave of successes, including dramatic reductions in disease burden in some areas and complete elimination of malaria in others [4]. These success stories have stimulated a renewed sense of optimism about prospects for global eradication [5].

If it is to succeed, this nascent drive toward country-by-country elimination and possible eventual worldwide eradication of malaria will require powerful new tools, importantly including malaria vaccines that produce protective immune responses that surpass those acquired through natural exposure to malaria [6]. Successful global or regional campaigns to eradicate smallpox, polio, and measles have all relied on vaccines. Those for yellow fever, hookworm, and yaws – and for malaria – which have relied on non-vaccine measures such as vector control or drug treatment have all failed [7]. While the odds of successful global malaria eradication would be very long even with an ideal malaria vaccine, they are virtually nil without one. In the meantime, even a modestly effective vaccine could substantially reduce the continuing heavy burden of malaria-attributable disease (247 million annual cases) and death (881,000 annual deaths) [2].

After nearly a century of malaria vaccine research, today, one modestly effective vaccine based on a parasite surface protein is being tested in a large Phase 3 trial in hopes of licensure within a few years [8]. Many other vaccine candidates have fallen short and been abandoned before reaching this stage, although several are in early stages of clinical development. The chief reasons that it has taken this long to get this far are that malaria parasites replicate and propagate through an extremely complex life cycle involving vertebrate hosts and an insect vector, and that they have evolved a repertoire of mechanisms for evading both natural and vaccine-induced immunity.

This review focuses on key themes that have emerged over 75 years of malaria vaccine research and development and on a few key examples of malaria vaccines that have reached the stage of testing for efficacy in clinical trials in humans. Several recent review articles listed after the primary bibliography explore malaria immunology and preclinical vaccine development in more detail.

The Malaria Life Cycle

Malaria is a potentially fatal parasitic disease transmitted to humans and other vertebrate animals by mosquitoes. Four species of *Plasmodium* cause malaria disease primarily in humans: *P. falciparum, P. vivax, P. ovale* and *P. malariae.* A fifth, *P. knowlesi*, infects mainly nonhuman primates but was recently found also to infect and sicken humans [9], and hundreds of other malaria species infect other mammals, reptiles, and birds. Because it is responsible for most severe malaria and deaths, *P. falciparum* has been the target of most vaccine development efforts and is the main focus of this review. However, in many parts of the world, *P. vivax* is the predominant species, and it causes more severe malaria than has sometimes been appreciated [10]. A vivax malaria vaccine would be highly beneficial, especially if it interrupted transmission.

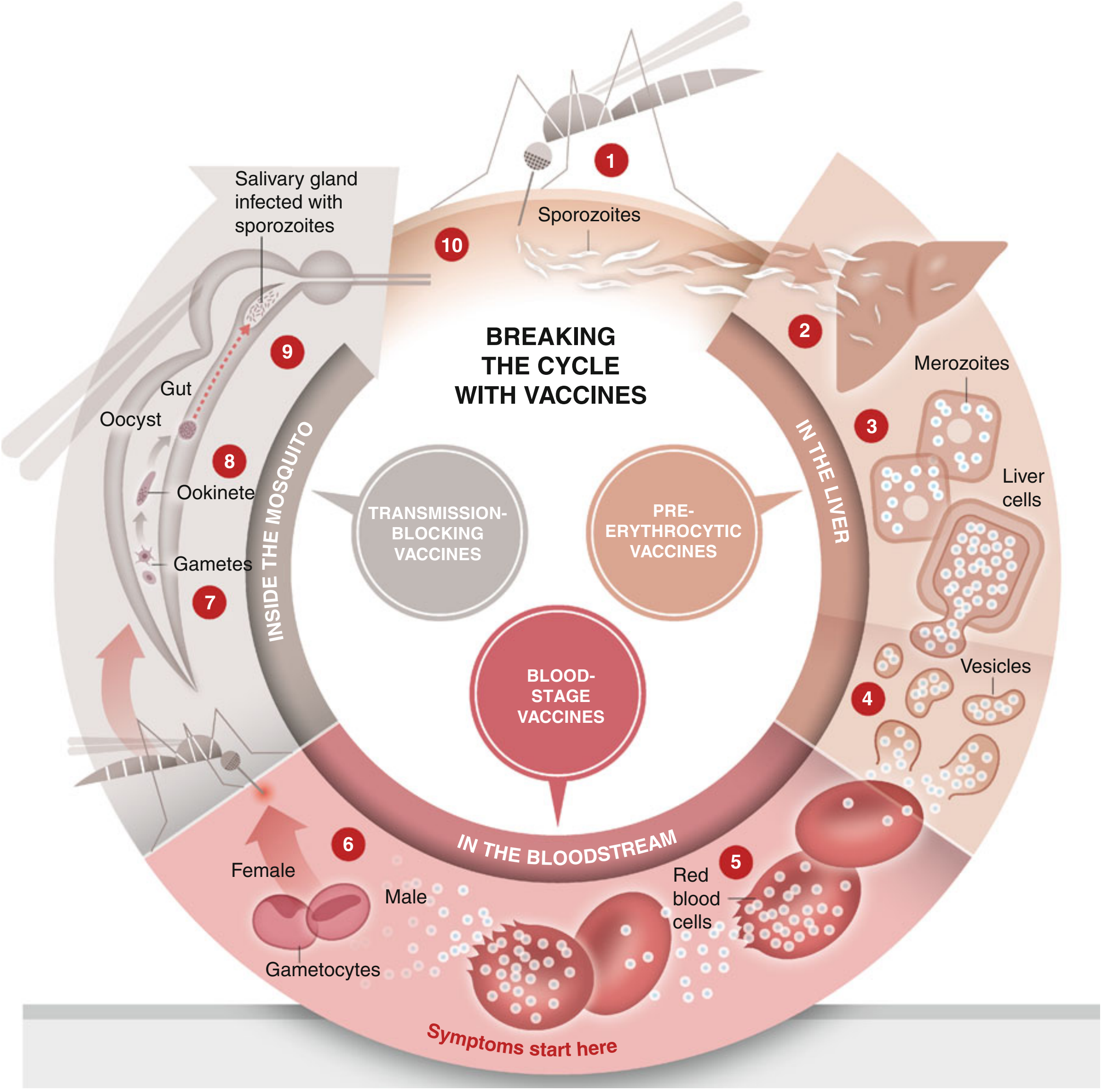
The malaria life cycle begins when the female *Anopheles* mosquito injects sporozoites from her salivary gland into the skin as she takes a blood meal (Fig. 1). The worm-like sporozoites – about 7 μ in length, or as long as an erythrocyte is wide – invade liver hepatocytes, each sporozoite multiplying over several days into tens of thousands of tiny (about 1 μ in diameter) merozoites packed into a single infected hepatocyte. These pre-erythrocytic stages cause no clinical signs or symptoms. A highly efficacious pre-erythrocytic vaccine would thus completely block infection, preventing parasites from reaching the blood and causing disease, and also preventing transmission.

Rupturing hepatocytes release showers of merozoites into the circulation, initiating the blood stage of malaria infection that is responsible for disease. Merozoites quickly invade erythrocytes and undergo asexual multiplication, dividing, growing, bursting from the erythrocyte, and re-invading in a periodic pattern with each cycle lasting 2 days (3 days in the case of *P. malariae*), until interrupted by host immunity, drug treatment, or death. Malaria vaccines that target the blood stage are thought of as anti-disease vaccines and would be expected to prevent or reduce clinical illness but would not prevent infection.

Some blood-stage parasites develop into male and female gametocytes. These sexual forms are taken up by the mosquito vector during a blood meal, mate to form a brief diploid stage, and then develop through further haploid stages and migrate from the gut to the salivary glands (Fig. 1). Each mating pair of gametocytes yields up to 1,000 infectious sporozoites, which are injected into the host to complete the transmission cycle. Vaccines targeting the sexual stages would prevent neither infection nor disease in the vaccinated individual and are thought of as transmission-blocking vaccines. Highly efficacious pre-erythrocytic or blood-stage vaccines that prevent sexual reproduction would also block transmission, so the term “transmission-blocking” need not refer exclusively to vaccines against sexual or mosquito stages of the parasite. The term “vaccines that interrupt transmission” (VIMT) encompasses all vaccines that interrupt transmission, whatever stage they target [11].

Pathogenesis and Disease

While semi-immune individuals can be chronically infected with malaria and experience no symptoms of illness, malaria infection of a nonimmune person usually causes an acute illness characterized by fever, chills, aches, and other flu-like symptoms. In a minority of cases, for reasons that are not well understood, more severe illness can develop. The clinical syndrome of falciparum malaria originates with changes in the infected erythrocytes. After they invade, blood-stage *P. falciparum* parasites effectively hijack the host cell and its machinery, expressing their own proteins on the surface of the host erythrocyte. Highly variant protein receptors called *P. falciparum* erythrocyte membrane proteins (PfEMP1) are encoded by a large, diverse family of up to 60 *var* genes in each parasite genome [12, 13]. PfEMP1 are expressed on the surface of infected red blood cells in clumps known as knobs, which are responsible for adherence of parasitized erythrocytes to the vascular endothelium, resulting in sequestration in tissue blood vessels [14].



Malaria Vaccines. Figure 1

Life cycle of malaria and stages targeted by vaccines. (Source: PATH – Malaria Vaccine Initiative). 1. Malaria infection begins when an infected female *Anopheles* mosquito bites a person, injecting *Plasmodium* parasites, in the form of sporozoites, into the bloodstream. 2. The sporozoites pass quickly into the human liver. 3. The sporozoites multiply asexually in the liver cells over the next 7–10 days, causing no symptoms. 4. The parasites, in the form of merozoites, burst from the liver cells. 5. In the bloodstream, the merozoites invade red blood cells (erythrocytes) and multiply again until the cells burst. Then, they invade more erythrocytes. This cycle is repeated, causing fever each time parasites break free and invade blood cells. 6. Some of the infected blood cells leave the cycle of asexual multiplication. Instead of replicating, the merozoites in these cells develop into sexual forms of the parasite, called gametocytes, that circulate in the bloodstream. 7. When a mosquito bites an infected human, it ingests the gametocytes, which develop further into mature sex cells called gametes. 8. The gametes develop into actively moving ookinetes that burrow into the mosquito's midgut wall and form oocysts. 9. Inside the oocyst, thousands of active sporozoites develop. The oocyst eventually bursts, releasing sporozoites that travel to the mosquito salivary glands. 10. The cycle of human infection begins again when the mosquito bites another person

The ability of falciparum malaria to sequester plays a critical role in disease severity – the other human malarias do not appear to sequester and, therefore, are not associated with most of the severe manifestations seen with falciparum malaria [15]. Infected red blood cells cytoadhere and sequester in the microcirculatory compartments of organs, most notably in the brain and placenta, leading to disease, and most abundantly in the spleen, causing splenomegaly. Sequestered infected red blood cells not only interfere with microcirculatory blood flow but also hide outside the reach of host defense mechanisms. Infected red blood cells lose their deformability and compromise blood flow in small capillaries and venules [16].

The presence of variant surface antigens leads to immune responses that appear both to harm the human host as well as to lead to the eventual development of protective immunity. Immunity to severe malaria develops rapidly, after only a few infections [17], possibly due to antibody responses that protect against a relatively conserved subset of PfEMP1 variants that are associated with severe malaria. In contrast, the slow acquisition of immune protection against uncomplicated malaria over years of repeated exposure to malaria is thought to represent the accumulation of protective immune responses to a repertoire of diverse antigens, probably including both PfEMP1 and the surface proteins that are the targets of most vaccine candidates [18].

Epidemiology

The epidemiology of malaria is determined primarily by the patterns and intensity of malaria transmission, which in turn drives the prevalence of malaria infection and the incidence of different forms of malaria disease. In low-transmission settings with unstable malaria, there is a potential for epidemic disease when transmission recurs or increases as a result of reintroduction to a population not recently exposed to malaria or to changing climactic or environmental conditions that favor contact between humans and malaria-transmitting *Anopheles* mosquitoes. Outbreaks can occur when malaria-naïve populations such as transmigrants, miners, or soldiers are exposed to malaria, causing high rates of disease [19].

Depending largely on the degree of host immunity, the manifestations of malaria infection can range from completely asymptomatic parasitemia, to mild disease that can be treated on an outpatient basis with oral drugs, to acute catastrophic life-threatening illness requiring intensive care. Very young infants are thought to be protected from malaria disease by maternal antibodies and persistent hemoglobin F. While they may be infected congenitally or by mosquito bites in the first days or weeks of life, infants do not experience clinical disease until they are a few months old. Following this brief period of relative insusceptibility in early infancy, protective immunity against malaria disease is acquired through repeated exposure and is therefore related to transmission intensity.

Where malaria transmission is moderate (average of one or more infected mosquito bites per month) or high (two or more infected mosquito bites per week; up to more than one per day in some areas), the risk period for death from malaria is highest in infants and young children who are in the process of developing acquired immunity. In a typical moderate- or high-transmission setting in sub-Saharan Africa, most severe malaria is experienced by children aged less than 5 years; children aged up to 10–12 years experience frequent episodes of uncomplicated malaria; and older teenagers and adults, while still often infected, rarely experience symptoms of malaria illness. Severe anemia is more frequent in the youngest infants, while cerebral malaria tends to peak in children aged 3–4 years who have experienced previous malaria episodes, suggesting that an overly exuberant immune response contributes to the pathogenesis of cerebral malaria.

In contrast, in low-transmission settings, persons of all ages have a similar risk of infection and uncomplicated malaria, most who are infected become sick, and the risk for severe malaria persists throughout life. Semi-immune adults, although they remain susceptible to asymptomatic parasitemia, are protected against clinical malaria disease, rarely becoming ill even when persistently infected. This protective immunity is lost after a few years in the absence of exposure. Acquired immunity is also diminished in pregnancy, in that women pregnant with their first child are susceptible to severe *P. falciparum* disease from placental malaria because they lack immunity to placenta-specific cytoadherence proteins. As placental immunity develops in subsequent pregnancies, there is a reduced risk of adverse effects of malaria in pregnancy on the mother and fetus [20].

Based on these epidemiological patterns, the primary populations targeted for malaria vaccines are infants and young children in areas of moderate and high transmission who bear the greatest burden of disease and death, and women of childbearing age in these same areas. Malaria-naïve travelers and military troops would also benefit from a malaria vaccine. As more countries move toward malaria elimination and global eradication is considered [21], the general population of malaria-endemic areas may be vaccinated to drive down transmission [6, 11].

Immunity

Insights into malaria immunity come not only from studies in various animal models including birds, rodents, and nonhuman primates but also from important studies in humans, including classic passive transfer experiments [22, 23] and early studies of malaria therapy for neurosyphilis [24, 25], as well as immuno-epidemiological studies that try to identify correlates of clinical protection [26, 27]. Both humoral and cellular factors contribute to acquired immune protection against malaria. Broadly speaking, cellular immune responses are thought to be more important in controlling the pre-erythrocytic stages of malaria infection [28], and antibodies are thought to block erythrocyte invasion to suppress blood-stage infection [22, 26]. For these reasons, cellular immune responses are typically emphasized in the development of pre-erythrocytic vaccines and antibody responses in the development of blood-stage vaccines. However, despite nearly 100 years of human and animal research, the basis of protective immunity against malaria is poorly understood, and no specific immune response has been established as an essential correlate of clinical protection, complicating malaria vaccine development.

New genomic tools have the potential to improve understanding of malaria immunity and may aid in vaccine development. For example, while it has long been known that there is some degree of strain specificity to malaria immunity, the discovery of large families of genes encoding highly variable surface antigens that mediate cytoadherence and immune evasion [14] has led to models for explaining the slow acquisition of protective immunity as a process of building up a repertoire of variant-specific immune responses until protection is in place against the full range of locally prevalent variants. As next-generation sequencing technologies improve their ability to generate sequence from clinical samples and to assemble genomes and map variant sequences to a reference genome, genomic epidemiology studies that relate parasite genotypes to clinical risk and allele-specific immune responses will permit testing of the hypotheses generated by such models. In another new approach, high-density protein arrays permit serological profiling of large numbers of serum samples against thousands of recombinant malaria proteins [29]. When this protein array was used to identify *P. falciparum* proteins that were differentially recognized by the sera of children who were resistant to clinical malaria, several previously unknown antigens were identified as possibly being important in acquired immunity, providing possible new vaccine targets [30].

In addition to acquired immunity, several host genetic factors offer some degree of protection against malaria, generally not by preventing infection but by reducing the risk of clinical illness or severe disease. Sickle cell trait [31] and other hemoglobinopathies [32–34] are more prevalent in populations at risk of malaria because they afford protection against clinical malaria. Various other human genetic polymorphisms associated with the host immune response [35] and with host–parasite binding [36] have also been correlated with susceptibility to clinical malaria in genetic association studies.

Early Malaria Vaccines

In 1880, Charles Louis Alphonse Laveran, a 33-year-old French Army doctor working in Algeria, discovered motile worm-like parasites, later understood to be exflagellating male gametes, in the blood of a feverish soldier [37]. Seven years later, Ronald Ross established that malaria parasites were transmitted to birds by the bites of infected mosquitoes [38]. The tremendous public health benefit that would be provided by an effective malaria vaccine was quickly appreciated, and from the 1930s to the 1970s, malaria vaccine researchers used primarily birds (including ducklings, canaries, chickens, and turkeys) and occasionally monkeys as model systems, and inactivated or killed whole parasites or parasite extracts as vaccines, often accompanied by immune-boosting adjuvant systems. This wave of vaccine development research crested in the late 1940s as World War II ended and attention shifted to the global campaign to eradicate malaria using drugs and anti-vector methods, and only resurged in the late 1960s when it became apparent that eradication was not possible with existing tools.

Early work focused on attenuated whole-parasite vaccines. Working in India, Russell and Mohan protected chickens from mosquito challenge with *P. gallinaceum* by immunizing them with sporozoites inactivated by ultraviolet light [39]. In 1945, the Hungarian-American immunologist Jules Freund (of Freund's complete adjuvant fame) and colleagues reported that they had successfully protected ducks against intravenous challenge with the avian malaria *P. lophurae* by immunizing them with formalin-inactivated malaria-infected blood cells and an adjuvant system consisting of a lanolin-like substance, paraffin oil, and killed tubercle bacilli [40]. They used a similar vaccine formulation to protect rhesus monkeys against *P. knowlesi* challenge [41]. These pioneering studies demonstrated two important principles that remain highly relevant to contemporary malaria vaccine development efforts, namely that good protective efficacy can be achieved with whole-organism vaccines and that strong immune-boosting adjuvants can achieve levels of protection that match or exceed those acquired through repeated natural exposure.

In a thoughtful and prescient paper published in 1943 describing protection of ducklings when a bacterial toxin adjuvant was added to a killed blood-stage vaccine, Henry Jacobs briefly cited an abstract presented by W. B. Redmond at the 1939 annual meeting of the American Society of Parasitologists, writing that "Redmond. . .noted some protection against bird malaria when he vaccinated with irradiated parasites" [42]. Redmond's abstract described using a frozen killed vaccine but made no mention of inactivation by irradiation [43]. Unfortunately, Redmond never published this work, but one can speculate that he may have described using some form of a radiation-attenuated *P. lophurae* vaccine preparation in his presentation at the 1939 meeting, anticipating subsequent work using this approach, including attenuation by irradiation of both blood stages [44, 45] and later and more famously of sporozoites [46, 47].

Columbia University scientists, who were aware of "inconclusive" earlier studies reported in the German medical literature, described in 1946 their own unsuccessful attempts to protect humans against intravenous challenge with *P. vivax* using a vaccine consisting of formalin-treated and freeze-thawed blood containing 65–150 million blood-stage *P. vivax* parasites [48]. No adjuvant was used in these earliest human studies, which may partially explain the disappointing results.

Several of these historical threads came together in a series of major breakthroughs in the late 1960s and early 1970s. First, Ruth Nussenzweig and Jerome Vanderberg at New York University reported in 1967 that intravenous immunization with irradiated *P. berghei* sporozoites could protect mice against subsequent intravenous challenge with viable sporozoites [46]. This advance was quickly translated into human trials of radiation-attenuated *P. falciparum* sporozoites delivered by the bites of infected, irradiated mosquitoes [49, 50]. In these and subsequent malaria challenge trials, 90% of volunteers who were immunized with radiation-attenuated sporozoites by receiving at least 1,000 infected bites over several sessions were fully protected against infection [51]. All unvaccinated volunteers acquired malaria from the bites of non-irradiated mosquitoes. These pioneering studies, first done by University of Maryland investigators in prisoners [52], provided proof that humans could be protected against infection with deadly *P. falciparum* through immunization, and spurred the identification of specific sporozoite proteins that could serve as antigens for subunit vaccines, as described in the following sections.

At around this time, two major advances ushered in the modern era of malaria vaccine development: the advent of molecular biology and the ability to clone, sequence, and express parasite genes in heterologous expression systems such as bacteria and yeast, and the development of methods for growing *P. falciparum* parasites in continuous in vitro culture [53, 54], providing a reliable and reproducible source of malaria parasites, proteins, and genes.

Obstacles to Malaria Vaccines

Why is there still no licensed malaria vaccine after 75 years of vaccine development research and evaluation of more than 70 vaccine candidates [55] in preclinical and clinical testing? Obstacles slowing progress toward an effective malaria vaccine include the size and complexity of the parasite, its genetic diversity, the efficiency of its amplification, the incomplete and temporary nature of naturally acquired immunity, and the fact that in addition to providing protection, immune responses also contribute to pathogenesis. Furthermore, parasite material must generally be obtained from infected hosts or mosquitoes – only one species, *P. falciparum*, can be grown in continuous culture. Finally, validated immune correlates of protection are lacking, so candidate vaccines can only be downselected by conducting costly efficacy trials in humans.

The *P. falciparum* genome has about 23 million bases of DNA organized into 14 chromosomes and about 5,000 genes [56]. This is orders of magnitude larger than the genomes of the viruses and bacteria to which vaccines have been successfully developed. This complexity and the large number of gene products provide the means and materials needed for the highly complex life cycle stages in vertebrate hosts and mosquitoes (Fig. 1). Moreover, mutation during mitotic reproduction in the haploid liver and blood stages and genetic recombination during the diploid sexual reproductive stages in the mosquito result in extensive genetic diversity that is driven by selection pressure from the immune system, as well as by drugs and, when they are deployed, potentially by vaccines [57]. All of this complexity and diversity greatly complicates the choice of candidate antigens for vaccine development.

At least 18 different forms of one leading blood-stage antigen [58] and more than 200 variants of another [59] have been documented in a single African village. If vaccines targeting these antigens generate immune responses that are insufficiently cross-protective, vaccines based on just one or two genetic variants are unlikely to be broadly efficacious [57]. To date, the choices of which variants of target antigens to include in malaria vaccines have not been made in consideration of the frequencies of these variants in natural populations. Careful molecular epidemiological studies are beginning to pinpoint which of the many polymorphisms in some of these antigens are the most important determinants of strain-specific natural immunity [58, 59], and this approach may help inform the design of polyvalent or chimeric vaccines that protect against diverse parasite strains [57].

Immunization with whole parasites of a given life cycle stage or with stage-specific proteins typically protects against only that life cycle stage, hence the notion of vaccines that prevent infection, disease, or transmission by targeting the different stages. While a highly efficacious pre-erythrocytic vaccine would prevent not only infection, but by doing so also prevent disease as well as transmission [6], even a single surviving sporozoite could theoretically result in a full-blown infection, severe disease, and transmission. This is because of the parasite's ability to multiply rapidly – one sporozoite gives rise to tens of thousands of merozoites emerging from the liver about a week and a half after a mosquito bite, and each merozoite multiplies roughly tenfold during the 48-h blood cycle, quickly resulting in billions of parasites circulating in the body. In reality, the rate at which parasites amplify their numbers is determined not just by the parasite's maximum reproductive capacity but also by host defenses and other factors. The ability of a partially efficacious pre-erythrocytic vaccine to reduce the risk of clinical malaria illness [60] supports the idea that there is some benefit from slowing the rate of parasite reproduction short of complete prevention of blood-stage infection – a so-called leaky vaccine, perhaps better thought of as an injectable bednet.

Most successful vaccines prevent infection or illness with pathogens that naturally result in strong and long-lasting immune protection after a single exposure. As described above, the naturally acquired protective immunity to malaria is hard won and short lived. An effective malaria vaccine would need to produce stronger immune responses more quickly than those that develop even under intense continuous natural exposure to malaria, and a vaccine intended to prevent infection will need to surpass natural immunity, which gradually protects against clinical illness but does not completely prevent infection.

Because the host immune response contributes to malaria pathogenesis, a vaccine could theoretically increase the risk of harmful inflammatory responses to subsequent infection, especially for vaccines directed against the blood stages that are responsible for pathology. One blood-stage malaria vaccine based on the *P. falciparum* merozoite surface protein 1 (MSP1) protected monkeys against lethal infection but then resulted in life-threatening anemia following subsequent exposure to malaria [61]. No such post-vaccination anemia has been observed in malaria-exposed adults [62, 63] or children [64, 65] in African trials of a different MSP1 malaria vaccine. A Phase 2 trial of a blood-stage malaria vaccine based on a different antigen, the apical membrane antigen 1 (AMA1), reported a possible increased risk of anemia in vaccinated malaria-exposed children in an unplanned post-hoc analysis [66]. No increased risk of anemia has been seen in trials with a more highly immunogenic AMA1 vaccine tested in similar populations [67, 68]. Vigilance for untoward inflammatory responses to malaria vaccines or to post-vaccination malaria infection will continue to be an important aspect of clinical malaria vaccine development, especially for blood-stage vaccines.

Pre-erythrocytic Vaccines

The vaccine furthest along in clinical development, RTS,S/AS01, targets the pre-erythrocytic circumsporozoite protein (CSP) of *P. falciparum*. The gene encoding CSP was the first *P. falciparum* gene cloned [69], and as the major surface protein coating sporozoites, CSP was immediately of great interest as a vaccine candidate. Thought to be important in sporozoite development and motility [70], CSP contains a central repeat region that elicits antibody responses [71], flanked on each side by non-repetitive regions containing T-cell epitopes [69]. Antibodies directed against the central repeat region cause the protein coat to slough off and block invasion of hepatocytes, suggesting that vaccine-induced antibodies might prevent infection [72]. Early synthetic CSP vaccines based on the repeat region using aluminum hydroxide as an adjuvant were poorly immunogenic, although a few individuals who achieved high antibody titers were protected against experimental challenge with homologous sporozoites [73, 74].

Seroepidemiological studies failed to find an association between anti-CSP antibodies and protection against infection [75], however, and the addition of adjuvants that produced higher antibody levels did not result in improved efficacy [76], leading the developers of RTS,S to include the flanking regions containing T-cell epitopes in subsequent versions of the vaccine. The final form of RTS,S is comprised of the central repeat region (R) and T-cell epitopes (T) using the hepatitis B surface antigen (S) as a carrier matrix, and co-expressed in *Saccharomyces cerevisiae* with additional S, hence “RTS,S.” The clinical development of RTS,S has included progressive improvements in adjuvant systems, resulting in improved efficacy both in experimental challenges [77–79] and in clinical trials in malaria-exposed adults [80] and children [60, 81]. The current formulation includes the liposomal-based Adjuvant System AS01, which contains the immunostimulants monophosphoryl lipid A and QS21, a saponin derivative extracted from the bark of the South American soap bark tree *Quillaja saponaria*. The development of RTS,S supports the notion that strong adjuvants are a necessary component for efficacious subunit protein malaria vaccines.

RTS,S/AS01 (and its immediate forebears with oil-in-water versions of the AS0-adjuvant system) was the first malaria vaccine to demonstrate meaningful levels of clinical protection in field trials. Trials in children and infants who are naturally exposed to malaria have demonstrated efficacy against clinical disease in the range of 30–56% and up to 66% against infection, and a good record of safety and tolerability [60, 81, 82]. A large Phase 2 trial in 1,465 Mozambican children showed 26% efficacy against all malaria episodes over nearly 4 years, 32% against first or only episode of clinical malaria, and 38% in preventing severe clinical episodes [83]. After 45 months, the prevalence of parasitemia was significantly lower in vaccines compared to the control group (12% vs 19%). The magnitude of the protective effect after nearly 4 years was thus modest, but importantly, there was no evidence of a post-immunization “rebound” effect – a theoretical concern that the vaccine might interfere with the natural acquisition of protective immunity.

Based on these demonstrations of a level of efficacy that is well below levels of protection expected for vaccines against other common pathogens but rare good news for malaria vaccines, RTS,S/AS01 is currently being evaluated in a large Phase 3 trial of

16,000 children and infants in seven African countries. The cost-effectiveness of licensing and deploying a malaria vaccine with efficacy in the range of 25–50% is debated, but where the malaria burden remains high, as in much of sub-Saharan Africa, such a vaccine is likely to be sought and used. Strategies being investigated to improve on the efficacy of RTS,S/AS01 include adding antigens tested with the same adjuvant system to create a multistage, multi-antigen RTS,S-based vaccine [84] and priming with an adenovirus expressing CSP before boosting with RTS,S/AS01 [85].

Several other pre-erythrocytic vaccine candidates have progressed through various stages of preclinical and early clinical development [55], including recombinant subunit protein vaccines as well as DNA vaccines and viral vectored vaccines [81]. Even though some of these have generated seemingly good humoral and especially cellular immune responses when formulated with strong adjuvants [86], protective efficacy has not been achieved in clinical testing in humans. Prime-boost approaches using DNA vaccines or viral vectors have also resulted in improved immunogenicity including cell-mediated responses in some participants and, in some cases, in measurable delays in time to infection in experimental sporozoite challenge trials [87]. Although prime-boost vaccine strategies have failed to demonstrate meaningful protection in published clinical efficacy trials, recent unpublished reports are more promising, with about 25% sterile protection provided by DNA prime and viral vector boost using both CSP and the blood-stage antigen AMA1 (T. Richie, personal communication). Efforts to improve these approaches to get more consistent cellular immune responses and higher levels of protection may be hampered by variability in host responses to vaccination. Other novel approaches such as a self-assembling polypeptide nanoparticle CSP vaccine are showing promise in preclinical testing [88]. Mining of genomic and proteomic data has led to the identification of new pre-erythrocytic vaccine candidate proteins, some of which have shown promising results in early preclinical testing in animal models [89].

Blood-Stage Vaccines

Most blood-stage malaria vaccine candidates are based on antigens that coat the surface of the invasive merozoites and/or that are involved with the process of erythrocyte invasion, in hopes of generating antibodies that block invasion and curtail parasite replication in the blood, reducing the risk or severity of clinical illness. The merozoite surface protein 1, or MSP1, was the first and best characterized of many proteins on the merozoite surface that are being targeted for vaccine development. MSP1 undergoes cleavage into four fragments that remain on the merozoite surface as a complex. Before erythrocyte invasion, the entire MSP1 complex is shed except for the C-terminal 19-kDa fragment ($MSP1_{19}$), which remains on the surface as the merozoite enters the erythrocyte [90]. Naturally acquired antibodies to $MSP1_{19}$ inhibit erythrocyte invasion and are associated with protection from clinical malaria in field studies [91, 92], supporting its potential as a vaccine candidate. Studies of recombinant MSP1 vaccines in monkeys were encouraging [93]. An $MSP1_{19}$-based vaccine on the same adjuvant platform as RTS,S produced antibodies in Malian adults that recognized MSP1 from diverse strains of *P. falciparum* [62], but had no protective efficacy against clinical malaria in Kenyan children [65]. Comparison of the degree of homology with the vaccine strain of MSP1 sequences in the infections experienced by children in the vaccine and control groups will clarify the extent to which the genetic diversity of MSP1 accounted for this lack of efficacy.

The apical membrane antigen 1, or AMA1, resides in the apical complex of the merozoite [94] before being processed and moving to the surface as the merozoite is released from the infected erythrocyte [95], where it is thought to play a role in erythrocyte invasion [96]. Proteomic studies have shown that AMA1 is also expressed in the sporozoite stage [97], suggesting that it may play a similar role in hepatocyte invasion. People living in malaria-endemic areas produce antibodies to AMA1 that can inhibit erythrocyte invasion in vitro [98] and that are associated with protection in field studies [99]. Studies in animal models show strain specificity in the inhibitory activity of anti-AMA1 antibodies [98], and these results have been corroborated by subsequent allelic exchange experiments [100, 101].

Sequencing of the gene encoding AMA1 in samples from a single Malian village identified more than 200 unique AMA1 variants in about 500 *P. falciparum* infections [59], raising the daunting prospect that a 200-valent AMA1 vaccine might be required to achieve broad protective efficacy. However, molecular epidemiological analyses showed that a group of just eight polymorphic amino acids lying adjacent to the presumed erythrocyte-binding site on AMA1 were responsible for strain-specific naturally acquired immunity, suggesting that a vaccine comprised of as few as ten "serotypes" of AMA1 might be sufficient to protect against 80% of unique variants.

Two AMA1 vaccines have reached the stage of efficacy trials in humans. A bivalent vaccine with two different forms of AMA1 adjuvanted with aluminum hydroxide failed to provide any protection against parasitemia or clinical malaria [66], and molecular analyses of pre- and post-immunization infections turned up no evidence of strain-specific efficacy or selection of non-vaccine variants [102]. A monovalent AMA1 vaccine formulated with the same adjuvant system as that used with RTS,S was more highly immunogenic [68] in a similar population of African children. Although this vaccine did not prevent infection after experimental sporozoite challenge [103] and showed marginal overall efficacy against clinical malaria, it demonstrated strong strain-specific efficacy, reducing the risk of clinical malaria caused by parasites with AMA1 homologous to the vaccine strain by more than 60% [104]. This encouraging result suggests that it may be possible to develop a more broadly efficacious multivalent or chimeric next-generation AMA1 vaccine, and efforts are being made to do this [105–107].

The *P. falciparum* proteins that are expressed on the surface of infected erythrocytes and that mediate cytoadherence and immune evasion and contribute to pathogenesis would seem to be attractive candidates for anti-disease vaccines. These PfEMP1 proteins are encoded by a large family of diverse *var* genes [14] with up to 60 variants in each parasite genome. Designing a vaccine that would be broadly protective against such an extraordinarily polymorphic target is likely to be very difficult. One possible approach to overcome this difficulty may be the identification of conserved epitopes that are nevertheless immunogenic [108]. Because a single PfEMP1 that is somewhat less polymorphic, VAR2CSA, mediates cytoadherence in placental malaria, prospects for a PfEMP1-based pregnancy malaria vaccine may be better, and research toward this goal is underway [109].

Multistage, multi-antigen vaccines that include blood-stage components are discussed in a subsequent section.

Transmission-Blocking Vaccines

Transmission-blocking vaccines are vaccines that are specifically intended to block transmission by targeting molecules that are unique to gametocytes, the male and female forms that are taken up during a blood meal and that mate in the mosquito midgut, or that target subsequent mosquito stages. Antibodies directed against such targets are capable of blocking the development of mosquito stages, thus interrupting transmission [110]. In a rare example of a vaccine designed to target multiple species, a vaccine based on mosquito-stage proteins in both *P. falciparum* and *P. vivax* was recently shown to produce dose-dependent antibody-mediated transmission-blocking activity [111]. However, the vaccine, which was formulated with the powerful adjuvant Montanide ISA 51, was unacceptably reactogenic. With the recent renewed call for global malaria eradication [5, 21], transmission-blocking vaccines will be increasingly emphasized. In one novel and promising approach, vaccines that target mosquito molecules are being contemplated in hopes of avoiding selection pressure within the host that favors "vaccine-resistant" parasites [112].

Although not traditionally thought of as transmission-blocking vaccines, highly efficacious pre-erythrocytic vaccines that provide sterile immunity would also interrupt transmission. Because they would completely prevent infection, such vaccines would also have the important added benefit of preventing disease caused by the blood stages. In the context of malaria elimination, a highly efficacious pre-erythrocytic vaccine would thus be the product development target for "vaccines that interrupt transmission" [11].

Multistage, Multi-antigen Vaccines

Compared to the viral and bacterial human pathogens for which effective vaccines exist, malaria parasites are big and complex and elicit equally complex and multifaceted immune responses. In retrospect, it may not be surprising that so many candidate vaccines that target just a single variant of a single antigen have failed to demonstrate clinical efficacy, especially against heterologous natural challenge. Several attempts have been made to improve on the efficacy of single-antigen vaccines by developing multistage, multi-antigen vaccines. One of the earliest was SPf66, a synthetic vaccine consisting of peptides derived from the blood-stage antigen MSP1 linked by the central repeat of the pre-erythrocytic antigen CSP. While reports of efficacy in initial trials in South America generated great excitement [113], subsequent studies in Africa and Asia showed no significant protective efficacy [114–116]. In another approach that yielded disappointing results, vaccinia virus was used as a vector to express seven *P. falciparum* genes, but both immunogenicity and efficacy were limited [117]. Attempts to develop DNA vaccines with from two to as many as 15 *P. falciparum* genes [118] were likewise unsuccessful.

Based on the modest efficacy of RTS,S against clinical malaria and evidence that a blood-stage vaccine using a similar adjuvant system can produce strain-specific efficacy [104], it is reasonable to believe that it may be possible to construct a multistage, multi-antigen recombinant protein that improves on the efficacy of RTS,S [84]. However, it seems not unlikely that vaccines that target 2, 5, or even 15 of the 5,000 gene products will still fall short of the high levels of protection seen with radiation-attenuated whole-organism vaccines when delivered through the bites of infected mosquitoes [51].

Whole-Organism Vaccines

Even though live attenuated vaccines were the earliest and remain some of the best vaccines against other pathogens, and even though birds and monkeys had been protected by live attenuated malaria parasites in the earliest vaccine studies [40, 41], the protection seen with irradiated sporozoites in the early 1970s [49] was interpreted not as a direct path to a malaria vaccine but as proof that a vaccine was possible and as justification for the ensuing decades of research aimed at identifying the "right" vaccine antigen or heterologous expression system. The limited success of these Sisyphean research efforts led to reevaluation of the dogma that it would be impossible to manufacture an attenuated sporozoite vaccine in mosquitoes [51]. A radiation-attenuated, metabolically active, non-replicating sporozoite vaccine has been manufactured in and purified from aseptically raised mosquitoes [47] and was recently evaluated for safety and efficacy in an experimental sporozoite challenge trial in humans. The goal was to administer by needle essentially the same immunogen – albeit aseptic, purified, and cryopreserved – that had previously demonstrated 90% protective efficacy in the form of sterilizing immunity when delivered by the bites of at least 1,000 irradiated infected mosquitoes. When administered by intradermal or subcutaneous routes, the sporozoite vaccine did not have significant protective efficacy [119]. Contemporaneous studies show that the vaccine sporozoites are highly immunogenic in monkeys when administered intravenously but not subcutaneously [119]. The likeliest explanation for this result is thought to be that sporozoites delivered into the skin by needle injection in a comparatively large volume of fluid were unable to reach the liver with efficiency approaching that of sporozoites injected either by a mosquito probing for small blood vessels or intravenously.

The potential for this approach to yield a highly efficacious pre-erythrocytic whole-organism vaccine remains well worth pursuing. Efforts are underway to improve delivery methods to more closely approximate the probing mosquito's efficient delivery of attenuated sporozoites into the circulation, starting with the intravenous route that is clearly superior in animal models. Whole-parasite vaccine experiments in the 1940s showed that adjuvants could boost protection, and this will likely be tried with the sporozoite vaccine. Genetic attenuation of sporozoites as an alternative to radiation is also being explored [120]. In a similar "back to the future" paradigm shift, attenuated whole-parasite blood-stage vaccines are now also back on the table [121, 122] more than 60 years after this approach was shown to work in monkeys.

Future Directions

While about 30 of the more than 100 countries with malaria transmission are actively trying to eliminate malaria, it is generally agreed that global malaria eradication is not possible without either global economic development to the levels that permitted malaria elimination in the United States, Europe, and the former Soviet Union or new tools such as a highly efficacious malaria vaccine that interrupts transmission [11]. A powerfully adjuvanted pre-erythrocytic single-antigen recombinant protein vaccine, RTS,S/AS01, significantly reduces the clinical burden of malaria in African children [60, 81], but it does not prevent infection, and, somewhat surprisingly, its effect on transmission, if any, has not been reported. While the results of a field trial of a similarly adjuvanted blood-stage vaccine [104] as well as combination of RTS,S with a viral vector in a prime-boost regimen [85] provide a rationale for pursuing improvements on RTS,S and similar vaccines, the dire public health need for a highly efficacious malaria vaccine calls for more than incremental improvements. Whole-parasite vaccines may be the radically different new (and yet very old) approach that is needed. Challenges remain to produce and deliver such a vaccine, but none that are insurmountable.

Research from the 1940s and many more recent studies point to the importance of strong adjuvants for subunit as well as for whole-organism vaccines, and wide access to immunogenic and safe adjuvants will be important for accelerating malaria vaccine development. Researchers are focusing now on the painfully elusive goal of achieving a high degree of efficacy, but the ideal vaccine would be not only safe and efficacious but also thermostable and protective for a long period of time after a single immunization – characteristics that will not be easy to achieve but which might be possible through pharmaceutical technologies such as controlled release formulations.

All predictions of when a malaria vaccine will be available have been overly optimistic, but barring unforeseen setbacks, a vaccine that substantially reduces the malaria burden should be licensed within just a few years. This will be a magnificent accomplishment that will mark not the end of the road for malaria vaccine development but a critical milestone on the path leading toward the malaria vaccine that the world needs.

Acknowledgments

The author is supported by the National Institute of Allergy and Infectious Diseases of the U.S. National Institutes of Health, by the Doris Duke Charitable Foundation, and by the Howard Hughes Medical Institute and wishes to thank Kirsten Lyke, Kavita Gandhi, Matthew Laurens, Mark Travassos, Thomas Richie, and Judith Epstein for critical reading of the manuscript and to acknowledge the U.S. Agency for International Development, the Walter Reed Army Institute of Research, GlaxoSmithKline Biologicals, Sanaria Inc., the U.S. Military Malaria Vaccine Program – Naval Medical Research Center, and the Malaria Research and Training Center of the University of Bamako, Mali, for collaboration on malaria vaccine trials.

Bibliography

Primary Literature

1. Garnham PCC (1966) Malaria parasites and other haemosporidia. Blackwell, Oxford
2. World Health Organization (2008) The global malaria action plan for a malaria free world. World Health Organization, Geneva
3. Rieckmann KH (2006) The chequered history of malaria control: are new and better tools the ultimate answer? Ann Trop Med Parasitol 100:647–662. doi:10.1179/136485906X112185
4. WHO (2007) United Arab Emirates certified malaria-free. Wkly Epidemiol Rec 82:30–32
5. Roberts L, Enserink M (2007) Malaria. Did they really say. . .eradication? Science 318:1544–1545
6. Plowe CV, Alonso P, Hoffman SL (2009) The potential role of vaccines in the elimination of falciparum malaria and the eventual eradication of malaria. J Infect Dis 200:1646–1649. doi:10.1086/646613
7. Henderson DA (1999) Lessons from the eradication campaigns. Vaccine 17(Suppl 3):S53–S55, S0264410X99002935 [pii]
8. Casares S, Brumeanu TD, Richie TL (2010) The RTS,S malaria vaccine. Vaccine 28:4880–4894. doi:10.1016/j.vaccine.2010.05.033, S0264-410X(10)00721–8 [pii]
9. Cox-Singh J, Davis TM, Lee KS, Shamsul SS, Matusop A, Ratnam S, Rahman HA, Conway DJ, Singh B (2008) *Plasmodium knowlesi* malaria in humans is widely distributed and potentially life threatening. Clin Infect Dis 46:165–171. doi:10.1086/524888
10. Price RN, Tjitra E, Guerra CA, Yeung S, White NJ, Anstey NM (2007) Vivax malaria: neglected and not benign. Am J Trop Med Hyg 77:79–87, 77/6_Suppl/79 [pii]
11. The malERA Consultative Group on Vaccines (2010) A research agenda for malaria eradication: vaccines. PLoS Med 8:e1000398
12. Baruch DI, Pasloske BL, Singh HB, Bi X, Ma XC, Feldman M, Taraschi TF, Howard RJ (1995) Cloning the *P. falciparum* gene encoding PfEMP1, a malarial variant antigen and adherence receptor on the surface of parasitized human erythrocytes. Cell 82:77–87
13. Su X, Heatwole VM, Wertheimer SP, Guinet F, Herrfeldt JA, Peterson DS, Ravetch JV, Wellems TE (1995) The large diverse gene family *var* encodes proteins involved in cytoadherence and antigenic variation of *Plasmodium falciparum*-infected erythrocytes. Cell 82:89–100

14. Smith JD, Chitnis CE, Craig AG, Roberts DJ, Hudson-Taylor DE, Peterson DS, Pinches R, Newbold CI, Miller LH (1995) Switches in expression of *Plasmodium falciparum* var genes correlate with changes in antigenic and cytoadherent phenotypes of infected erythrocytes. Cell 82:101–110
15. Miller LH, Baruch DI, Marsh K, Doumbo OK (2002) The pathogenic basis of malaria. Nature 415:673–679
16. Nash GB, O'Brien E, Gordon-Smith EC, Dormandy JA (1989) Abnormalities in the mechanical properties of red blood cells caused by *Plasmodium falciparum*. Blood 74:855–861
17. Gupta S, Snow RW, Donnelly CA, Marsh K, Newbold C (1999) Immunity to non-cerebral severe malaria is acquired after one or two infections. Nat Med 5:340–343
18. Hviid L (2005) Naturally acquired immunity to *Plasmodium falciparum* malaria in Africa. Acta Trop 95:270–275. doi:10.1016/j.actatropica.2005.06.012, S0001-706X(05)00165-8
19. Baird JK, Jones TR, Danudirgo EW, Annis BA, Bangs MJ, Basri H, Purnomo MS (1991) Age-dependent acquired protection against *Plasmodium falciparum* in people having two years exposure to hyperendemic malaria. Am J Trop Med Hyg 45:65–76
20. Duffy PE (2007) Plasmodium in the placenta: parasites, parity, protection, prevention and possibly preeclampsia. Parasitology 134:1877–1881. doi:10.1017/S0031182007000170, S0031182007000170 [pii]
21. Tanner M, de Savigny D (2008) Malaria eradication back on the table. Bull World Health Organ 86:82, S0042-96862008000200002 [pii]
22. Cohen S, McGregor IA, Carrington S (1961) Gamma-globulin and acquired immunity to human malaria. Nature 192:733–737
23. McGregor IA, Carrington S, Cohen S (1963) Treatment of East African *P. falciparum* malaria with West African human gamma-globulin. Trans R Soc Trop Med Hyg 50:170–175
24. Collins WE, Jeffery GM (1999) A retrospective examination of sporozoite- and trophozoite-induced infections with *Plasmodium falciparum* in patients previously infected with heterologous species of Plasmodium: effect on development of parasitologic and clinical immunity. Am J Trop Med Hyg 61:36–43
25. Collins WE, Jeffery GM (1999) A retrospective examination of secondary sporozoite- and trophozoite-induced infections with *Plasmodium falciparum*: development of parasitologic and clinical immunity following secondary infection. Am J Trop Med Hyg 61:20–35
26. Thomas AW, Trape JF, Rogier C, Goncalves A, Rosario VE, Narum DL (1994) High prevalence of natural antibodies against *Plasmodium falciparum* 83-kilodalton apical membrane antigen (PF83/AMA-1) as detected by capture-enzyme-linked immunosorbent assay using full-length baculovirus recombinant PF83/AMA-1. Am J Trop Med Hyg 51:730–740
27. Egan AF, Chappel JA, Burghaus PA, Morris JS, McBride JS, Holder AA, Kaslow DC, Riley EM (1995) Serum antibodies from malaria-exposed people recognize conserved epitopes formed by the two epidermal growth factor motifs of MSP1(19), the carboxy-terminal fragment of the major merozoite surface protein of *Plasmodium falciparum*. Infect Immun 63:456–466
28. Schofield L (1989) T cell immunity to malaria sporozoites. Exp Parasitol 68:357–364
29. Doolan DL, Mu Y, Unal B, Sundaresh S, Hirst S, Valdez C, Randall A, Molina D, Liang X, Freilich DA, Oloo JA, Blair PL, Aguiar JC, Baldi P, Davies DH, Felgner PL (2008) Profiling humoral immune responses to *P. falciparum* infection with protein microarrays. Proteomics 8:4680–4694
30. Crompton PD, Kayala MA, Traore B, Kayentao K, Ongoiba A, Weiss GE, Molina DM, Burk CR, Waisberg M, Jasinskas A, Tan X, Doumbo S, Doumtabe D, Kone Y, Narum DL, Liang X, Doumbo OK, Miller LH, Doolan DL, Baldi P, Felgner PL, Pierce SK (2010) A prospective analysis of the Ab response to *Plasmodium falciparum* before and after a malaria season by protein microarray. Proc Natl Acad Sci USA 107:6958–6963. doi:10.1073/pnas.1001323107, 1001323107 [pii]
31. Allison AC (1954) Protection afforded by sickle-cell trait against subtertian malarial infection. Br Med J 1:290–294
32. Kay AC, Kuhl W, Prchal J, Beutler E (1992) The origin of glucose-6-phosphate-dehydrogenase (G6PD) polymorphisms in African-Americans. Am J Hum Genet 50:394–398
33. Agarwal A, Guindo A, Cissoko Y, Taylor JG, Coulibaly D, Kone A, Kayentao K, Djimde A, Plowe CV, Doumbo O, Wellems TE, Diallo D (2000) Hemoglobin C associated with protection from severe malaria in the Dogon of Mali, a West African population with a low prevalence of hemoglobin S. Blood 96:2358–2363
34. Flint J, Hill AV, Bowden DK, Oppenheimer SJ, Sill PR, Serjeantson SW, Bana-Koiri J, Bhatia K, Alpers MP, Boyce AJ (1986) High frequencies of alpha-thalassaemia are the result of natural selection by malaria. Nature 321:744–750
35. Hill AV, Allsopp CE, Kwiatkowski D, Anstey NM, Twumasi P, Rowe PA, Bennett S, Brewster D, McMichael AJ, Greenwood BM (1991) Common west African HLA antigens are associated with protection from severe malaria. Nature 352:595–600
36. Miller LH, Mason SJ, Clyde DF, McGinniss MH (1976) The resistance factor to *Plasmodium vivax* in blacks. The Duffy- blood-group genotype, FyFy. N Engl J Med 295:302–304
37. Bruce-Chwatt LJ (1981) Alphonse Laveran's discovery 100 years ago and today's global fight against malaria. J R Soc Med 74:531–536
38. Ross R (1897) Observations on a condition necessary to the transformation of the malaria crescent. Br Med J 1:251–255
39. Russell PF, Mohan BN (1942) The immunization of fowls against mosquito-borne *Plamodium gallinaceum* by injections of serum and of inactivated homologous sporozoites. J Exp Med 76:477–495
40. Freund J, Sommer HE, Walter AW (1945) Immunization against malaria: vaccination of ducks with killed parasites incorporated with adjuvants. Science 102:200–202
41. Freund J, Thomson KJ, Sommer HE, Walter AW, Schenkein EL (1945) Immunization of rhesus monkeys against malarial infection (*P. knowlesi*) with killed parasites and adjuvants. Science 102:202–204. doi:10.1126/science.102.2643.202, 102/2643/202 [pii]
42. Jacobs HR (1943) Immunization against malaria. Increased protection by vaccination of ducklings with saline-insoluble

residues of *Plasmodium lophurae* mixed with a bacterial toxin. Am J Trop Med Hyg s1-23:597–606

43. Redmond WB (1939) Immunization of birds to malaria by vaccination. J Parasitol 25:28–29
44. Bennison BE, Coatney GR (1949) Effects of X-irradiation on *Plasmodium gallinaceum* and *Plasmodium lophurae* infections in young chicks. J Natl Malar Soc 8:280–289
45. Ceithaml J, Evans EA Jr (1946) The biochemistry of the malaria parasite; the in vitro effects of x-rays upon *Plasmodium gallinaceum*. J Infect Dis 78:190–197
46. Nussenzweig RS, Vanderberg J, Most H, Orton C (1967) Protective immunity produced by the injection of x-irradiated sporozoites of *Plasmodium berghei*. Nature 216:160–162
47. Hoffman SL, Billingsley PF, James E, Richman A, Loyevsky M, Li T, Chakravarty S, Gunasekera A, Li M, Stafford R, Ahumada A, Epstein JE, Sedegah M, Reyes S, Richie TL, Lyke KE, Edelman R, Laurens M, Plowe CV, Sim BK (2010) Development of a metabolically active, non-replicating sporozoite vaccine to prevent *Plasmodium falciparum* malaria. Hum Vaccin 6:97–106, 10396 [pii]
48. Heidelberger M, Prout C, Hindle JA, Rose AS (1946) Studies in human malaria III. An attempt at vaccination of paretics against blood-borne infection with *P vivax*. J Immunol 53: 109–112
49. Clyde DF, Most H, McCarthy VC, Vanderberg JP (1973) Immunization of man against sporozoite-induced falciparum malaria. Am J Med Sci 266:169–177
50. Rieckmann KH, Carson PE, Beaudoin RL, Cassells JS, Sell KW (1974) Letter: sporozoite induced immunity in man against an Ethiopian strain of *Plasmodium falciparum*. Trans R Soc Trop Med Hyg 68:258–259
51. Luke TC, Hoffman SL (2003) Rationale and plans for developing a non-replicating, metabolically active, radiation-attenuated *Plasmodium falciparum* sporozoite vaccine. J Exp Biol 206:3803–3808
52. Vanderberg JP (2009) Reflections on early malaria vaccine studies, the first successful human malaria vaccination, and beyond. Vaccine 27:2–9. doi:10.1016/j.vaccine.2008.10.028, S0264-410X(08)01414-X [pii]
53. Trager W, Jensen JB (1976) Human malaria parasites in continuous culture. Science 193:673–675
54. Haynes JD, Diggs CL, Hines FA, Desjardins RE (1976) Culture of human malaria parasites *Plasmodium falciparum*. Nature 263:767–769
55. World Health Organization (2010) Initiative for vaccine research: malaria vaccines. WHO, Geneva
56. Gardner MJ, Hall N, Fung E, White O, Berriman M, Hyman RW, Carlton JM, Pain A, Nelson KE, Bowman S, Paulsen IT, James K, Eisen JA, Rutherford K, Salzberg SL, Craig A, Kyes S, Chan MS, Nene V, Shallom SJ, Suh B, Peterson J, Angiuoli S, Pertea M, Allen J, Selengut J, Haft D, Mather MW, Vaidya AB, Martin DM, Fairlamb AH, Fraunholz MJ, Roos DS, Ralph SA, McFadden GI, Cummings LM, Subramanian GM, Mungall C, Venter JC, Carucci DJ, Hoffman SL, Newbold C, Davis RW, Fraser CM, Barrell B (2002) Genome sequence of the human malaria parasite *Plasmodium falciparum*. Nature 419:498–511
57. Takala SL, Plowe CV (2009) Genetic diversity and malaria vaccine design, testing and efficacy: preventing and overcoming 'vaccine resistant malaria'. Parasite Immunol 31:560–573. doi:10.1111/j.1365–3024.2009.01138.x, PIM1138 [pii]
58. Takala SL, Coulibaly D, Thera MA, Dicko A, Smith DL, Guindo AB, Kone AK, Traore K, Ouattara A, Djimde AA, Sehdev PS, Lyke KE, Diallo DA, Doumbo OK, Plowe CV (2007) Dynamics of polymorphism in a malaria vaccine antigen at a vaccine-testing site in Mali. PLoS Med 4:e93
59. Takala SL, Coulibaly D, Thera MA, Batchelor AH, Cummings MP, Escalante AA, Ouattara A, Traore K, Niangaly A, Djimde AA, Doumbo OK, Plowe CV (2009) Extreme polymorphism in a vaccine antigen and risk of clinical malaria: implications for vaccine development. Sci Transl Med 1:2ra5
60. Alonso PL, Sacarlal J, Aponte JJ, Leach A, Macete E, Milman J, Mandomando I, Spiessens B, Guinovart C, Espasa M, Bassat Q, Aide P, Ofori-Anyinam O, Navia MM, Corachan S, Ceuppens M, Dubois MC, Demoitie MA, Dubovsky F, Menendez C, Tornieporth N, Ballou WR, Thompson R, Cohen J (2004) Efficacy of the RTS, S/AS02A vaccine against *Plasmodium falciparum* infection and disease in young African children: randomised controlled trial. Lancet 364:1411–1420
61. Egan AF, Blackman MJ, Kaslow DC (2000) Vaccine efficacy of recombinant *Plasmodium falciparum* merozoite surface protein 1 in malaria-naive, -exposed, and/or -rechallenged *Aotus vociferans* monkeys. Infect Immun 68:1418–1427
62. Thera MA, Doumbo OK, Coulibaly D, Diallo DA, Sagara I, Dicko A, Diemert DJ, Heppner DG Jr, Stewart VA, Angov E, Soisson L, Leach A, Tucker K, Lyke KE, Plowe CV (2006) Safety and allele-specific immunogenicity of a malaria vaccine in Malian adults: results of a phase I randomized trial. PLoS Clin Trials 1:e34. doi:10.1371/journal.pctr.0010034
63. Stoute JA, Gombe J, Withers MR, Siangla J, McKinney D, Onyango M, Cummings JF, Milman J, Tucker K, Soisson L, Stewart VA, Lyon JA, Angov E, Leach A, Cohen J, Kester KE, Ockenhouse CF, Holland CA, Diggs CL, Wittes J, Heppner DG Jr (2007) Phase 1 randomized double-blind safety and immunogenicity trial of *Plasmodium falciparum* malaria merozoite surface protein FMP1 vaccine, adjuvanted with AS02A, in adults in western Kenya. Vaccine 25:176–184
64. Withers MR, McKinney D, Ogutu BR, Waitumbi JN, Milman JB, Apollo OJ, Allen OG, Tucker K, Soisson LA, Diggs C, Leach A, Wittes J, Dubovsky F, Stewart VA, Remich SA, Cohen J, Ballou WR, Holland CA, Lyon JA, Angov E, Stoute JA, Martin SK, Heppner DG (2006) Safety and reactogenicity of an MSP-1 malaria vaccine candidate: a randomized Phase Ib dose-escalation trial in Kenyan children. PLoS Clin Trials 1:e32
65. Ogutu BR, Apollo OJ, McKinney D, Okoth W, Siangla J, Dubovsky F, Tucker K, Waitumbi JN, Diggs C, Wittes J, Malkin E, Leach A, Soisson LA, Milman JB, Otieno L, Holland CA, Polhemus M, Remich SA, Ockenhouse CF, Cohen J, Ballou WR, Martin SK, Angov E, Stewart VA, Lyon JA, Heppner DG, Withers MR (2009) Blood stage malaria vaccine eliciting high

antigen-specific antibody concentrations confers no protection to young children in Western Kenya. PLoS One 4:e4708

66. Sagara I, Dicko A, Ellis RD, Fay MP, Diawara SI, Assadou MH, Sissoko MS, Kone M, Diallo AI, Saye R, Guindo MA, Kante O, Niambele MB, Miura K, Mullen GE, Pierce M, Martin LB, Dolo A, Diallo DA, Doumbo OK, Miller LH, Saul A (2009) A randomized controlled phase 2 trial of the blood stage AMA1-C1/Alhydrogel malaria vaccine in children in Mali. Vaccine 27:3090–3098
67. Thera MA, Doumbo OK, Coulibaly D, Diallo DA, Kone AK, Guindo AB, Traore K, Dicko A, Sagara I, Sissoko MS, Baby M, Sissoko M, Diarra I, Niangaly A, Dolo A, Daou M, Diawara SI, Heppner DG, Stewart VA, Angov E, Bergmann-Leitner ES, Lanar DE, Dutta S, Soisson L, Diggs CL, Leach A, Owusu A, Dubois MC, Cohen J, Nixon JN, Gregson A, Takala SL, Lyke KE, Plowe CV (2008) Safety and immunogenicity of an AMA-1 malaria vaccine in Malian adults: results of a Phase 1 randomized controlled trial. PLoS One 3:e1465
68. Thera MA, Doumbo OK, Coulibaly D, Laurens MB, Kone AK, Guindo AB, Traore K, Sissoko M, Diallo DA, Diarra I, Kouriba B, Daou M, Dolo A, Baby M, Sissoko MS, Sagara I, Niangaly A, Traore I, Olotu A, Godeaux O, Leach A, Dubois MC, Ballou WR, Cohen J, Thompson D, Dube T, Soisson L, Diggs CL, Takala SL, Lyke KE, House B, Lanar DE, Dutta S, Heppner DG, Plowe CV (2010) Safety and immunogenicity of an AMA1 malaria vaccine in Malian children: results of a phase 1 randomized controlled trial. PLoS One 5:e9041. doi:10.1371/journal.pone.0009041
69. Dame JB, Williams JL, McCutchan TF, Weber JL, Wirtz RA, Hockmeyer WT, Maloy WL, Haynes JD, Schneider RD (1984) Structure of the gene encoding the immunodominant surface antigen on the sporozoite of the human malaria parasite *Plasmodium falciparum*. Science 225:593–599
70. Beier JC, Vanderburg JP (1998) Sporogonic development in the mosquito. In: Sherman IW (ed) Malaria parasite biology, pathogenesis, and protection. ASM Press, Washington, DC
71. Ballou WR, Rothbard J, Wirtz RA, Gordon DM, Williams JS, Gore RW, Schneider I, Hollingdale MR, Beaudoin RL, Maloy WL et al (1985) Immunogenicity of synthetic peptides from circumsporozoite protein of *Plasmodium falciparum*. Science 228:996–999
72. Hollingdale MR, Nardin EH, Tharavanij S, Schwartz AL, Nussenzweig RS (1984) Inhibition of entry of *Plasmodium falciparum* and *P. vivax* sporozoites into cultured cells; an in vitro assay of protective antibodies. J Immunol 132:909–913
73. Herrington DA, Clyde DF, Losonsky G, Cortesia M, Murphy JR, Davis J, Baqar S, Felix AM, Heimer EP, Gillessen D et al (1987) Safety and immunogenicity in man of a synthetic peptide malaria vaccine against *Plasmodium falciparum* sporozoites. Nature 328:257–259
74. Ballou WR, Hoffman SL, Sherwood JA, Hollingdale MR, Neva FA, Hockmeyer WT, Gordon DM, Schneider I, Wirtz RA, Young JF (1987) Safety and efficacy of a recombinant DNA *Plasmodium falciparum* sporozoite vaccine. Lancet 1:1277–1281
75. Hoffman SL, Oster CN, Plowe CV, Woollett GR, Beier JC, Chulay JD, Wirtz RA, Hollingdale MR, Mugambi M (1987) Naturally acquired antibodies to sporozoites do not prevent malaria: vaccine development implications. Science 237: 639–642
76. Sherwood JA, Copeland RS, Taylor KA, Abok K, Oloo AJ, Were JB, Strickland GT, Gordon DM, Ballou WR, Bales JDJ, Wirtz RA, Wittes J, Gross M, Que JU, Cryz SJ, Oster CN, Roberts CR, Sadoff JC (1996) *Plasmodium falciparum* circumsporozoite vaccine immunogenicity and efficacy trial with natural challenge quantitation in an area of endemic human malaria of Kenya. Vaccine 14:817–827
77. Gordon DM, McGovern TW, Krzych U, Cohen JC, Schneider I, LaChance R, Heppner DG, Yuan G, Hollingdale M, Slaoui M (1995) Safety, immunogenicity, and efficacy of a recombinantly produced *Plasmodium falciparum* circumsporozoite protein-hepatitis B surface antigen subunit vaccine. J Infect Dis 171:1576–1585
78. Stoute JA, Slaoui M, Heppner DG, Momin P, Kester KE, Desmons P, Wellde BT, Garcon N, Krzych U, Marchand M (1997) A preliminary evaluation of a recombinant circumsporozoite protein vaccine against *Plasmodium falciparum* malaria. RTS, S malaria vaccine evaluation group. N Engl J Med 336:86–91. doi:10.1056/NEJM199701093360202
79. Kester KE, McKinney DA, Tornieporth N, Ockenhouse CF, Heppner DG, Hall T, Krzych U, Delchambre M, Voss G, Dowler MG, Palensky J, Wittes J, Cohen J, Ballou WR (2001) Efficacy of recombinant circumsporozoite protein vaccine regimens against experimental *Plasmodium falciparum* malaria. J Infect Dis 183:640–647
80. Bojang KA, Milligan PJ, Pinder M, Vigneron L, Alloueche A, Kester KE, Ballou WR, Conway DJ, Reece WH, Gothard P, Yamuah L, Delchambre M, Voss G, Greenwood BM, Hill A, McAdam KP, Tornieporth N, Cohen JD, Doherty T (2001) Efficacy of RTS, S/AS02 malaria vaccine against *Plasmodium falciparum* infection in semi-immune adult men in The Gambia: a randomised trial. Lancet 358:1927–1934
81. Bejon P, Lusingu J, Olotu A, Leach A, Lievens M, Vekemans J, Mshamu S, Lang T, Gould J, Dubois MC, Demoitie MA, Stallaert JF, Vansadia P, Carter T, Njuguna P, Awuondo KO, Malabeja A, Abdul O, Gesase S, Mturi N, Drakeley CJ, Savarese B, Villafana T, Ballou WR, Cohen J, Riley EM, Lemnge MM, Marsh K, von Seidlein L (2008) Efficacy of RTS, S/AS01E vaccine against malaria in children 5 to 17 months of age. N Engl J Med 359:2521–2532. doi:10.1056/NEJMoa0807381, NEJMoa0807381 [pii]
82. Alonso PL, Sacarlal J, Aponte JJ, Leach A, Macete E, Aide P, Sigauque B, Milman J, Mandomando I, Bassat Q, Guinovart C, Espasa M, Corachan S, Lievens M, Navia MM, Dubois MC, Menendez C, Dubovsky F, Cohen J, Thompson R, Ballou WR (2005) Duration of protection with RTS, S/AS02A malaria vaccine in prevention of *Plasmodium falciparum* disease in Mozambican children: single-blind extended follow-up of a randomised controlled trial. Lancet 366: 2012–2018
83. Sacarlal J, Aide P, Aponte JJ, Renom M, Leach A, Mandomando I, Lievens M, Bassat Q, Lafuente S, Macete E, Vekemans J,

Guinovart C, Sigauque B, Sillman M, Milman J, Dubois MC, Demoitie MA, Thonnard J, Menendez C, Ballou WR, Cohen J, Alonso PL (2009) Long-term safety and efficacy of the RTS, S/AS02A malaria vaccine in Mozambican children. J Infect Dis 200:329–336. doi:10.1086/600119

84. Heppner DG Jr, Kester KE, Ockenhouse CF, Tornieporth N, Ofori O, Lyon JA, Stewart VA, Dubois P, Lanar DE, Krzych U, Moris P, Angov E, Cummings JF, Leach A, Hall BT, Dutta S, Schwenk R, Hillier C, Barbosa A, Ware LA, Nair L, Darko CA, Withers MR, Ogutu B, Polhemus ME, Fukuda M, Pichyangkul S, Gettyacamin M, Diggs C, Soisson L, Milman J, Dubois MC, Garcon N, Tucker K, Wittes J, Plowe CV, Thera MA, Duombo OK, Pau MG, Goudsmit J, Ballou WR, Cohen J (2005) Towards an RTS, S-based, multi-stage, multi-antigen vaccine against falciparum malaria: progress at the Walter Reed Army Institute of Research. Vaccine 23:2243–2250
85. Stewart VA, McGrath SM, Dubois PM, Pau MG, Mettens P, Shott J, Cobb M, Burge JR, Larson D, Ware LA, Demoitie MA, Weverling GJ, Bayat B, Custers JH, Dubois MC, Cohen J, Goudsmit J, Heppner DG Jr (2007) Priming with an adenovirus 35-circumsporozoite protein (CS) vaccine followed by RTS, S/AS01B boosting significantly improves immunogenicity to *Plasmodium falciparum* CS compared to that with either malaria vaccine alone. Infect Immun 75:2283–2290
86. Cummings JF, Spring MD, Schwenk RJ, Ockenhouse CF, Kester KE, Polhemus ME, Walsh DS, Yoon IK, Prosperi C, Juompan LY, Lanar DE, Krzych U, Hall BT, Ware LA, Stewart VA, Williams J, Dowler M, Nielsen RK, Hillier CJ, Giersing BK, Dubovsky F, Malkin E, Tucker K, Dubois MC, Cohen JD, Ballou WR, Heppner DG Jr (2010) Recombinant liver stage antigen-1 (LSA-1) formulated with AS01 or AS02 is safe, elicits high titer antibody and induces IFN-gamma/IL-2 CD4+ T cells but does not protect against experimental *Plasmodium falciparum* infection. Vaccine 28:5135–5144. doi:10.1016/j.vaccine.2009.08.046, S0264-410X(09)01231-6 [pii]
87. Hill AV, Reyes-Sandoval A, O'Hara G, Ewer K, Lawrie A, Goodman A, Nicosia A, Folgori A, Colloca S, Cortese R, Gilbert SC, Draper SJ (2010) Prime-boost vectored malaria vaccines: progress and prospects. Hum Vaccin 6:78–83, 10116 [pii]
88. Kaba SA, Brando C, Guo Q, Mittelholzer C, Raman S, Tropel D, Aebi U, Burkhard P, Lanar DE (2009) A nonadjuvanted polypeptide nanoparticle vaccine confers long-lasting protection against rodent malaria. J Immunol 183:7268–7277. doi:10.4049/jimmunol.0901957, jimmunol.0901957 [pii]
89. Bergmann-Leitner ES, Mease RM, de la Vega P, Savranskaya T, Polhemus M, Ockenhouse C, Angov E (2010) Immunization with pre-erythrocytic antigen CelTOS from *Plasmodium falciparum* elicits cross-species protection against heterologous challenge with *Plasmodium berghei*. PLoS One 5: e12294. doi:10.1371/journal.pone.0012294
90. Blackman MJ, Heidrich HG, Donachie S, McBride JS, Holder AA (1990) A single fragment of a malaria merozoite surface protein remains on the parasite during red cell invasion and is the target of invasion-inhibiting antibodies. J Exp Med 172: 379–382
91. Egan AF, Morris J, Barnish G, Allen S, Greenwood BM, Kaslow DC, Holder AA, Riley EM (1996) Clinical immunity to *Plasmodium falciparum* malaria is associated with serum antibodies to the 19-kDa C-terminal fragment of the merozoite surface antigen, PfMSP-1. J Infect Dis 173:765–9
92. Riley EM, Allen SJ, Wheeler JG, Blackman MJ, Bennett S, Takacs B, Schonfeld HJ, Holder AA, Greenwood BM (1992) Naturally acquired cellular and humoral immune responses to the major merozoite surface antigen (PfMSP1) of *Plasmodium falciparum* are associated with reduced malaria morbidity. Parasite Immunol 14:321–37
93. Stowers AW, Cioce V, Shimp RL, Lawson M, Hui G, Muratova O, Kaslow DC, Robinson R, Long CA, Miller LH (2001) Efficacy of two alternate vaccines based *on Plasmodium falciparum* merozoite surface protein 1 in an Aotus challenge trial. Infect Immun 69:1536–1546. doi:10.1128/IAI.69.3.1536-1546.2001
94. Peterson MG, Marshall VM, Smythe JA, Crewther PE, Lew A, Silva A, Anders RF, Kemp DJ (1989) Integral membrane protein located in the apical complex of *Plasmodium falciparum*. Mol Cell Biol 9:3151–3154
95. Narum DL, Thomas AW (1994) Differential localization of full-length and processed forms of PF83/AMA-1 an apical membrane antigen of *Plasmodium falciparum* merozoites. Mol Biochem Parasitol 67:59–68
96. Mitchell GH, Thomas AW, Margos G, Dluzewski AR, Bannister LH (2004) Apical membrane antigen 1, a major malaria vaccine candidate, mediates the close attachment of invasive merozoites to host red blood cells. Infect Immun 72:154–158
97. Florens L, Washburn MP, Raine JD, Anthony RM, Grainger M, Haynes JD, Moch JK, Muster N, Sacci JB, Tabb DL, Witney AA, Wolters D, Wu Y, Gardner MJ, Holder AA, Sinden RE, Yates JR, Carucci DJ (2002) A proteomic view of the *Plasmodium falciparum* life cycle. Nature 419:520–526
98. Hodder AN, Crewther PE, Anders RF (2001) Specificity of the protective antibody response to apical membrane antigen 1. Infect Immun 69:3286–94
99. Polley SD, Mwangi T, Kocken CH, Thomas AW, Dutta S, Lanar DE, Remarque E, Ross A, Williams TN, Mwambingu G, Lowe B, Conway DJ, Marsh K (2004) Human antibodies to recombinant protein constructs of *Plasmodium falciparum* Apical Membrane Antigen 1 (AMA1) and their associations with protection from malaria. Vaccine 23:718–728
100. Dutta S, Lee SY, Batchelor AH, Lanar DE (2007) Structural basis of antigenic escape of a malaria vaccine candidate. Proc Natl Acad Sci USA 104:12488–12493
101. Healer J, Murphy V, Hodder AN, Masciantonio R, Gemmill AW, Anders RF, Cowman AF, Batchelor A (2004) Allelic polymorphisms in apical membrane antigen-1 are responsible for evasion of antibody-mediated inhibition in *Plasmodium falciparum*. Mol Microbiol 52:159–168
102. Ouattara A, Mu J, Takala-Harrison S, Saye R, Sagara I, Dicko A, Niangaly A, Duan J, Ellis RD, Miller LH, Su XZ, Plowe CV, Doumbo OK (2010) Lack of allele-specific efficacy of

a bivalent AMA1 malaria vaccine. Malar J 9:175. doi:10.1186/1475-2875-9-175, 1475-2875-9-175 [pii]

103. Spring MD, Cummings JF, Ockenhouse CF, Dutta S, Reidler R, Angov E, Bergmann-Leitner E, Stewart VA, Bittner S, Juompan L, Kortepeter MG, Nielsen R, Krzych U, Tierney E, Ware LA, Dowler M, Hermsen CC, Sauerwein RW, de Vlas SJ, Ofori-Anyinam O, Lanar DE, Williams JL, Kester KE, Tucker K, Shi M, Malkin E, Long C, Diggs CL, Soisson L, Dubois MC, Ballou WR, Cohen J, Heppner DG Jr (2009) Phase 1/2a study of the malaria vaccine candidate apical membrane antigen-1 (AMA-1) administered in adjuvant system AS01B or AS02A. PLoS One 4:e5254
104. Thera MA, Doumbo OK, Coulibaly D, Laurens MB, Ouattara A, Kone AK, Guindo AB, Traore K, Traore I, Kouriba B, Diallo DA, Diarra I, Daou M, Dolo A, Tolo Y, Sissoko MS, Niangaly A, Sissoko M, Takala-Harrison S, Lyke KE, Wu Y, Blackwelder WC, Godeaux O, Vekemans J, Dubois M-C, Ballou WR, Cohen J, Thompson D, Dube T, Soisson L, Diggs CL, House B, Lanar DE, Dutta S, Heppner DG Jr., Plowe CV (2011) A field trial to assess a blood-stage malaria vaccine. N Engl J Med 365: 1004–13
105. Remarque EJ, Faber BW, Kocken CH, Thomas AW (2008) A diversity-covering approach to immunization with *Plasmodium falciparum* apical membrane antigen 1 induces broader allelic recognition and growth inhibition responses in rabbits. Infect Immun 76:2660–2670. doi:10.1128/IAI.00170-08, IAI.00170-08 [pii]
106. Remarque EJ, Faber BW, Kocken CH, Thomas AW (2008) Apical membrane antigen 1: a malaria vaccine candidate in review. Trends Parasitol 24:74–84. doi:10.1016/j.pt.2007.12.002, S1471-4922(07)00328-5 [pii]
107. Dutta S, Dlugosz LS, Clayton JW, Pool CD, Haynes JD, Gasser RA III, Batchelor AH (2010) Alanine mutagenesis of the primary antigenic escape residue cluster, c1, of apical membrane antigen 1. Infect Immun 78:661–671. doi:10.1128/IAI.00866-09, IAI.00866-09 [pii]
108. Klein MM, Gittis AG, Su HP, Makobongo MO, Moore JM, Singh S, Miller LH, Garboczi DN (2008) The cysteine-rich interdomain region from the highly variable *Plasmodium falciparum* erythrocyte membrane protein-1 exhibits a conserved structure. PLoS Pathog 4:e1000147. doi:10.1371/journal.ppat.1000147
109. Avril M, Cartwright MM, Hathaway MJ, Hommel M, Elliott SR, Williamson K, Narum DL, Duffy PE, Fried M, Beeson JG, Smith JD (2010) Immunization with VAR2CSA-DBL5 recombinant protein elicits broadly cross-reactive antibodies to placental *Plasmodium falciparum*-infected erythrocytes. Infect Immun 78:2248–2256. doi:10.1128/IAI.00410-09, IAI.00410-09 [pii]
110. Barr PJ, Green KM, Gibson HL, Bathurst IC, Quakyi IA, Kaslow DC (1991) Recombinant Pfs25 protein of *Plasmodium falciparum* elicits malaria transmission-blocking immunity in experimental animals. J Exp Med 174:1203–1208
111. Wu Y, Ellis RD, Shaffer D, Fontes E, Malkin EM, Mahanty S, Fay MP, Narum D, Rausch K, Miles AP, Aebig J, Orcutt A, Muratova O, Song G, Lambert L, Zhu D, Miura K, Long C, Saul A, Miller LH, Durbin AP (2008) Phase 1 trial of malaria transmission blocking vaccine candidates Pfs25 and Pvs25 formulated with montanide ISA 51. PLoS One 3:e2636. doi:10.1371/journal.pone.0002636
112. Dinglasan RR, Valenzuela JG, Azad AF (2005) Sugar epitopes as potential universal disease transmission blocking targets. Insect Biochem Mol Biol 35:1–10. doi:10.1016/j.ibmb.2004.09.005, S0965-1748(04)00159-6 [pii]
113. Valero MV, Amador LR, Galindo C, Figueroa J, Bello MS, Murillo LA, Mora AL, Patarroyo G, Rocha CL, Rojas M (1993) Vaccination with SPf66, a chemically synthesised vaccine, against *Plasmodium falciparum* malaria in Colombia. Lancet 341:705–710
114. Nosten F, Luxemburger C, Kyle DE, Ballou WR, Wittes J, Wah E, Chongsuphajaisiddhi T, Gordon DM, White NJ, Sadoff JC, Heppner DG (1996) Randomised double-blind placebo-controlled trial of SPf66 malaria vaccine in children in north-western Thailand. Shoklo SPf66 Malaria Vaccine Trial Group. Lancet 348:701–707
115. D'Alessandro U, Leach A, Drakeley CJ, Bennett S, Olaleye BO, Fegan GW, Jawara M, Langerock P, George MO, Targett GA (1995) Efficacy trial of malaria vaccine SPf66 in Gambian infants. Lancet 346:462–467
116. Alonso PL, Smith T, Schellenberg JR, Masanja H, Mwankusye S, Urassa H, Bastos de Azevedo I, Chongela J, Kobero S, Menendez C et al (1994) Randomised trial of efficacy of SPf66 vaccine against *Plasmodium falciparum* malaria in children in southern Tanzania. Lancet 344:1175–1181
117. Ockenhouse CF, Sun PF, Lanar DE, Wellde BT, Hall BT, Kester K, Stoute JA, Magill A, Krzych U, Farley L, Wirtz RA, Sadoff JC, Kaslow DC, Kumar S, Church LW, Crutcher JM, Wizel B, Hoffman S, Lalvani A, Hill AV, Tine JA, Guito KP, de Taisne C, Anders R, Ballou WR (1998) Phase I/IIa safety, immunogenicity, and efficacy trial of NYVAC-Pf7, a pox-vectored, multiantigen, multistage vaccine candidate for *Plasmodium falciparum* malaria. J Infect Dis 177:1664–1673
118. Kumar S, Epstein JE, Richie TL, Nkrumah FK, Soisson L, Carucci DJ, Hoffman SL (2002) A multilateral effort to develop DNA vaccines against falciparum malaria. Trends Parasitol 18:129–135, S1471492201022073 [pii]
119. Epstein JE, Tewari K, Lyke KE, Sim BK, Billingsley PF, Laurens MB, Gunasekera A, Chakravarty S, James ER, Sedegah M, Richman A, Velmurugan S, Reyes S, Li M, Tucker K, Ahumada A, Ruben AJ, Li T, Stafford R, Eappen AG, Tamminga C, Bennett JW, Ockenhouse CF, Murphy JR, Komisar J, Thomas N, Loyevsky M, Birkett A, Plowe CV, Loucq C, Edelman R, Richie TL, Seder RA, Hoffman SL (2011) Live attenuated malaria vaccine designed to protect through hepatic CD8+ T cell immunity. Science [Epub ahead of print]
120. VanBuskirk KM, O'Neill MT, de la Vega P, Maier AG, Krzych U, Williams J, Dowler MG, Sacci JB Jr, Kangwanrangsan N, Tsuboi T, Kneteman NM, Heppner DG Jr, Murdock BA, Mikolajczak SA, Aly AS, Cowman AF, Kappe SH (2009) Preerythrocytic, live-attenuated *Plasmodium falciparum* vaccine candidates by design. Proc Natl Acad Sci USA 106. doi:10.1073/pnas.0906387106, 13004–13009. 0906387106 [pii]

121. Pinzon-Charry A, McPhun V, Kienzle V, Hirunpetcharat C, Engwerda C, McCarthy J, Good MF (2010) Low doses of killed parasite in CpG elicit vigorous CD4+ T cell responses against blood-stage malaria in mice. J Clin Invest 120:2967–2978. doi:10.1172/JCI39222, 39222 [pii]
122. McCarthy JS, Good MF (2010) Whole parasite blood stage malaria vaccines: a convergence of evidence. Hum Vaccin 6:114–123, 10394 [pii]

Books and Reviews

Bruder JT, Angov E, Limbach KJ, Richie TL (2010) Molecular vaccines for malaria. Hum Vaccin 6:54–77

Desowitz RS (1991) The malaria capers: more tales of parasites and people, research and reality. Norton, New York

Dinglasan RR, Jacobs-Lorena M (2008) Flipping the paradigm on malaria transmission-blocking vaccines. Trends Parasitol 24:364–370

Hill AV, Reyes-Sandoval A, O'Hara G, Ewer K, Lawrie A, Goodman A, Nicosia A, Folgori A, Colloca S, Cortese R, Gilbert SC, Draper SJ (2010) Prime-boost vectored malaria vaccines: progress and prospects. Hum Vaccin 6:78–83

Hviid L (2010) The role of Plasmodium falciparum variant surface antigens in protective immunity and vaccine development. Hum Vaccin 6:84–89

Langhorne J, Ndungu FM, Sponaas AM, Marsh K (2008) Immunity to malaria: more questions than answers. Nat Immunol 9:725–732

Moorthy VS, Kieny MP (2010) Reducing empiricism in malaria vaccine design. Lancet Infect Dis 10:204–211

Richards JS, Beeson JG (2009) The future for blood-stage vaccines against malaria. Immunol Cell Biol 87:377–390

Shah S (2010) The fever: how malaria has ruled humankind for 500,000 years. Farrar, Straus and Giroux, New York

Sherwin IW (2009) The elusive malaria vaccine: miracle or mirage? ASM Press, Washington, DC

Vaughan AM, Wang R, Kappe SH (2010) Genetically engineered, attenuated whole-cell vaccine approaches for malaria. Hum Vaccin 6:107–113

Wykes M, Good MF (2007) A case for whole-parasite malaria vaccines. Int J Parasitol 37:705–712

Mariculture Systems, Integrated Land-Based

Muki Shpigel
Israel Oceanographic and Limnological Research,
National Center for Mariculture, Eilat, Israel

Article Outline

Glossary

Detritivores (also known as *saprophages*) They are heterotrophs that obtain nutrients by consuming detritus (decomposing organic matter).

Halophyte Salt-loving plants that can be grown at higher salinities than most traditional crop plants.

IMTA The Integrated Multi-Trophic Aquaculture System (IMTA) is an aquaculture practice in which excretions of one or more organisms are utilized by other cultured organisms from different trophic (nutritional) levels within the system.

Land-based and offshore mariculture systems Two methods of seawater aquaculture (mariculture); the former on land and the latter in the ocean.

Polyculture An aquaculture practice which involves culture of two or more species from the same or different trophic levels in the same water reservoir.

RAS Recirculated Aquaculture System (RAS) is an aquaculture practice for the rearing of aquatic organisms wherein 90% or more of the water is recycled within the system.

Sludge Solid/particulate waste that includes, among other components, feces, uneaten feed, algae and bacteria, which sinks to the bottom of aquaculture water reservoirs.

Definition of Subject

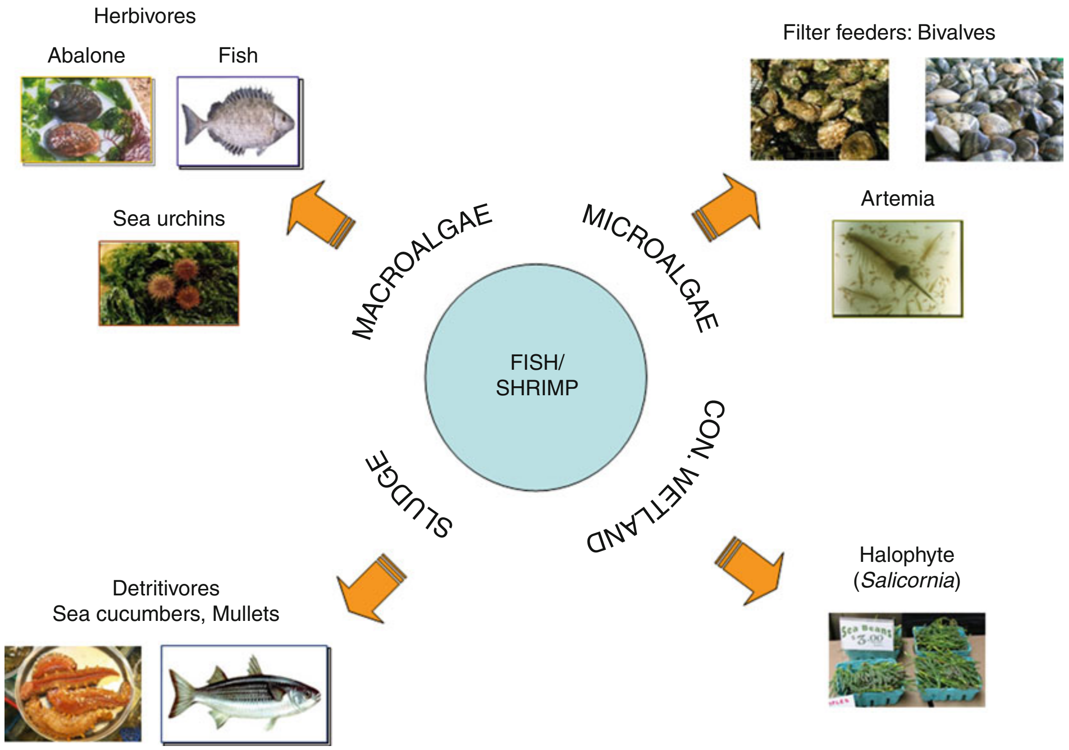
The Integrated Multi-Trophic Aquaculture System (IMTA) is an aquaculture practice in which excretions of one or more organisms are utilized by other cultured organisms from different trophic (nutritional) levels. IMTA systems are distinct from polyculture systems, which involve two or more species from the same or different trophic levels in the same water reservoir. In a typical IMTA, the various species are cultured in separate spatial entities, permitting intensification and optimization of production. The IMTA concept has been increasingly adopted in modern day aquaculture, including land-based (Fig. 1) [1–5] and offshore mariculture [6, 7].

In land-based IMTA systems, seawater is pumped from the sea to fish or shrimp ponds. A pelleted diet is the only source of nutrients for the animals in the system. Nutrient-rich effluent water from these ponds can take three directions: microalgae ponds, macroalgae ponds, and constructed wetlands with halophyte plants. The microalgae can be utilized by filter feeders such as *Artemia* or/and bivalves. The macroalgae can be utilized by macroalgivores such as abalone or sea urchins, and detritus can be utilized by detritivores such as mullets, sea cucumbers, or polychaete worms (Fig. 1).

Introduction

The concept of polyculture and IMTA systems is not new. Such systems of different species of fish, or combinations of invertebrates and fish, have been existing in ancient Egypt and China for thousands of years. Artificial enclosures or natural ponds in tidal zones were generally used. Extensive traditional IMTA and polyculture systems are still practiced today in various parts of Asia in fresh and salt water. Rice and fish are cultured together in China. Earthen ponds, in association with wild or agricultural plants, are used on a wide scale in fish and shrimp farming in China, Indonesia, Taiwan, Thailand, Japan, Vietnam, India, the Philippines, and Ecuador. In Europe, ducks, fish, and crayfish have been raised together in freshwater ponds. This type of extensive production has proven sustainable, because it utilizes organisms that feed on different levels of the food web, and maintains a clean environment.

The traditional IMTA and polyculture systems are more environmentally friendly than modern intensive mono-aquaculture systems. These systems utilize fewer resources and do not pollute surrounding waters with waste products, because they generally sustain relatively low stocking densities and do not employ fertilizers. Most of them rely on natural production of food.



Mariculture Systems, Integrated Land-Based. Figure 1
Schematic design of land-based IMTA systems (con. = constructed)

This concept has increasingly been adopted for modern aquaculture, including land-based and sea-cage mariculture. With dramatic increases in global human population, food demand, and overfishing problems, traditional extensive aquaculture cannot satisfy present demand, and much less so the projected future demand, for sea products.

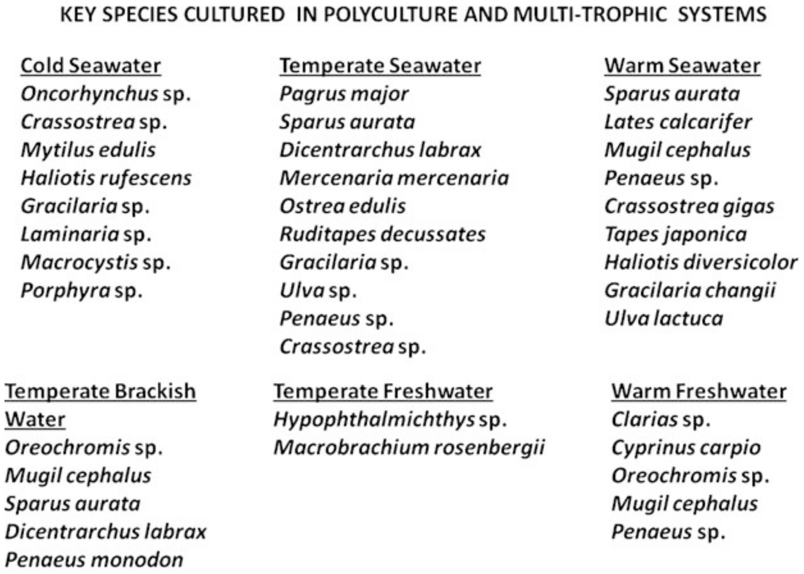
Modern intensive monoculture systems require high levels of resources and produce undesirable wastes. They are dedicated to a few expensive species and do not generate a large amount of food. Intensive aquaculture uses extensive amounts of resources such as water, feeds, fertilizers, chemicals, and energy, while discharging fecal material, uneaten feed, excretions, and drugs into the environment. In turn, this creates eutrophication of the water, has deleterious effects on marine life, increases the risks of antibiotic resistance in organisms, has an adverse effect on biodiversity, and contributes to habitat destruction. The economic success of intensive monoculture in sea cages or land-based facilities has much to do with the fact that, even today, pollution of the environment involves little or no monetary outlay or penalty for the growers. In most countries, aquaculture does not yet include the cost of effluent treatment. However, in the industrialized nations, this age is coming to a timely end and in Europe, there are already laws and regulations requiring effluent treatment and imposing fines for noncompliance. In some countries, this cost can be as high as €0.5–1 kg^{-1} feed, resulting in an expense of €250,000–350,000 per annum for medium-scale RAS (Recirculating Aquaculture System) farms (250 t/year). Awareness is growing among scientists, industry, the public, and politicians that technologies disregarding environmental impact are neither sustainable nor acceptable.

History

The development of modern land-based IMTA using extractive organisms such as shellfish, microalgae, and seaweeds began in the 1970s with the pioneer work of Goldman et al. [1] and Ryther et al. [2] in the treatment of household effluents. Phytoplankton was cultured in a mixture of domestic wastewater effluent and seawater, fed to suspension-feeder molluscs, and the dissolved remnants of nutrients in the final effluent were assimilated by seaweeds. As the food value of organisms grown on human waste effluents was questionable, adaptations of this principle to the treatment of intensive aquaculture effluents in both inland and coastal areas were proposed [8] and were followed by the integration into a system of carnivorous fish and abalone (e.g., [9]). The first practical and quantitative integrated land-based cultures of marine fish and shellfish, with phytoplankton as biofilter and food for shellfish, were constructed in Israel by Hughes-Games [10] and Gordin et al. [11]. A semi-intensive seabream and gray mullet pond system with silicate-rich green water, located on the coast of the Gulf of Eilat (Red Sea), supported dense populations of diatoms, excellent for feeding oysters [12, 13]. Later, the development of a practical intensive culture of bivalves in phytoplankton-rich effluents was described in a series of articles [3, 14–18]. Lefebvre et al. [19] showed that detritical waste from intensive fish farming can contribute to the growth of bivalves and reduce particulate matter in the water. Jones et al. [20], using the Sydney rock oyster *Saccostrea commercialis*, significantly reduced the concentration of suspended particulates including algae, bacteria, and inorganic particles in integrated systems.

Studies showing the performance of seaweed in land-based IMTA, initially at laboratory scale and later expanded to outdoor pilot scale, began to appear in the 1970s [21, 22]. The theoretical and practical principles of intensive large-scale land-based seaweed culture were studied and developed first at Woods Hole and later at Harbor Branch Oceanographic Institution in Florida – U.S.A. [8, 23–25]. The quantitative aspects of their functioning have been described [14, 16, 26–29]. Fish, abalone, and seaweed IMTA systems were studied by Shpigel et al. [30], Butterworth [31], and Nobre et al. [32]. The aspects of bioeconomics of land-based IMTA are described by Nobre et al. [32], Neori et al. [33], and Bunting and Shpigel [34].

Offshore IMTA system is a relatively new concept that started in the late nineties and is a modification of the land-based IMTA. In coastal integrated mariculture, shellfish and seaweed are cultured in proximity to cage fish culture [6, 7]. Kelp (brown algae) [35, 36] and red algae [37, 38] efficiently take up dissolved inorganic nitrogen excreted by the fish [39], so that seaweed production and quality are often higher in areas surrounding fish cages than elsewhere [6, 40–42].



KEY SPECIES CULTURED IN POLYCULTURE AND MULTI-TROPHIC SYSTEMS

Cold Seawater
Oncorhynchus sp.
Crassostrea sp.
Mytilus edulis
Haliotis rufescens
Gracilaria sp.
Laminaria sp.
Macrocystis sp.
Porphyra sp.

Temperate Seawater
Pagrus major
Sparus aurata
Dicentrarchus labrax
Mercenaria mercenaria
Ostrea edulis
Ruditapes decussates
Gracilaria sp.
Ulva sp.
Penaeus sp.
Crassostrea sp.

Warm Seawater
Sparus aurata
Lates calcarifer
Mugil cephalus
Penaeus sp.
Crassostrea gigas
Tapes japonica
Haliotis diversicolor
Gracilaria changii
Ulva lactuca

Temperate Brackish Water
Oreochromis sp.
Mugil cephalus
Sparus aurata
Dicentrarchus labrax
Penaeus monodon

Temperate Freshwater
Hypophthalmichthys sp.
Macrobrachium rosenbergii

Warm Freshwater
Clarias sp.
Cyprinus carpio
Oreochromis sp.
Mugil cephalus
Penaeus sp.

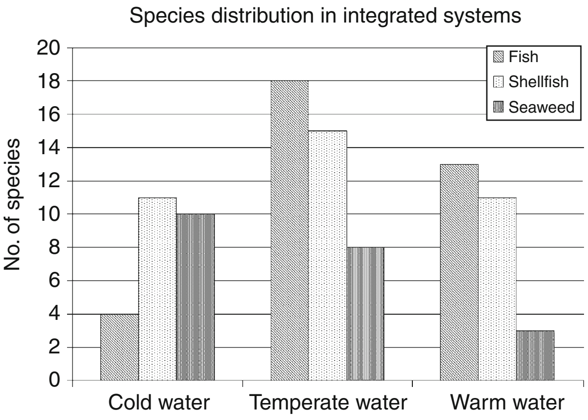
Mariculture Systems, Integrated Land-Based. Figure 2
Key species cultured in IMTA and polyculture systems for marine and freshwater environment

However, nutrient removal efficiency in offshore IMTA is still relatively low, ranging between 15% and 25% [43].

The concept of IMTA systems is generic and can be applied to cold, warm, and temperate waters, in intensive, semi-intensive, and extensive systems, in sea cages or land-based facilities, in fresh water in land-based facilities or lakes, and all of the above in closed, semi-closed, or flow-through systems.

In recent years, several enterprises and research facilities have begun setting up land-based IMTA; most of the systems are pilot scale or R&D facilities. The IMTA typically include two or three species. In most of the studies, seaweed and microalgae are used as biofilters for the dissolved nutrients (review by Neori et al. [33] and Soto [44]). A broad spectrum list of selected organisms being used in farms and in R&D is presented in Fig. 2. Key species in cold water are salmon, mussels, and the seaweeds *Gracilaria, Laminaria*, and *Porphyra*. For temperate and warm seawater, sea bream, sea bass, oysters, clams, and *Ulva lactuca* are the predominant cultured species (Fig. 2).

Over 200 species are currently the object of R&D projects and in commercial farms and research institutes around the world, in various climate conditions. A significant number of fish and shellfish are cultured in temperate water, and a relatively low number of fish and large number of seaweeds in cold-water climates (Fig. 3).



Mariculture Systems, Integrated Land-Based. Figure 3
Fish, shellfish, and seaweed species combination in IMTA systems in different bio-geographical regions around the world

Nutrient Budget in Land-Based IMTA

Protein in fish or shrimp feed is the most expensive component of nitrogen input into the IMTA systems.

In conventional cages or ponds, fish or shrimps assimilate only 20–30% of the nitrogen, while the rest is excreted into the water, mainly as dissolved ammonia, feces, and uneaten feed.

Two main practical approaches are emerging for handling the organic and nitrogenous wastes: bacterial dissimilation into gasses in "Recirculating Aquaculture Systems" (RAS), or plant assimilation into biomass (IMTA). Bacterial biofilters are dissimilative. Through a process of nitrification followed by denitrification, bacteria break down the organic pollutants into N_2 and CO_2 gasses. Bacterial biofilters are technically rather effective for aquaculture and allow significant water recirculation. However, the technology is relatively expensive, and not simple. Bacterial biofilter technologies are suitable for relatively small intensive land-based culture of lucrative organisms. There are no suggestions as to how such technologies can be integrated with large-scale, low cost fish or shrimp production. In addition, this system wastes expensive nitrogen by converting this valuable resource into gas, which is lost into the atmosphere.

Nutrient assimilation by other organisms is a more promising method of water treatment. In land-based IMTA ponds, seawater is pumped from the "nuclear species" (fish or shrimp) into the ponds/tanks of secondary organisms or macro-/microalgae. A pellet diet is the only source of nutrients for the primary animals in the system. Nutrient-rich effluent water from these ponds can take three directions: microalgae ponds, macroalgae ponds, or to irrigate halophyte crops (e.g., *Salicornia* sp.). The microalgae can be utilized by filter feeders such as artemia or bivalves. The macroalgae can be utilized by macroalgivores such as abalone, sea urchins, or herbivorous fish. Halophytes such as *Salicornia* can be used as a food product. The remaining detritus can be fed to detritivores such as mullets, sea cucumbers, or polychaete worms, singly or in combination.

Optimization of the IMTA is typically based on the highest value "nuclear" product at any given time. This "nuclear" product may be shifted according to climatic conditions and economic considerations. For example, in a fish-abalone-seaweed integrated system, abalone is the most valuable species, and the entire system is centered around this species. Abalone will be the first organism to receive the incoming water. From the abalone, the water will drain to the ammonia producers and from there to the biofilters.

The biological and chemical processes in the IMTA system should be balanced between nutrient production by the main organism and nutrient uptake capacity of the micro- and/or macroalgae and downstream by the micro- and macroalgivores. In such systems evaluated in Eilat, Israel, macro-and microalgae were able to assimilate 1–5 g N m^{-2} day^{-1}, while algivores and filter feeders assimilated 0.5–1 g N kg $(WW)^{-1}$ day^{-1} (Table 1 and references therein). However, there will be variation in nutrient uptake depending on season and climate, as algal biomass is influenced by day length (i.e., light hours), water temperature, and the nutrient levels in the water.

Mariculture Systems, Integrated Land-Based. Table 1 Assimilation rates of the uptake organisms in land-based IMTA in Eilat, Israel

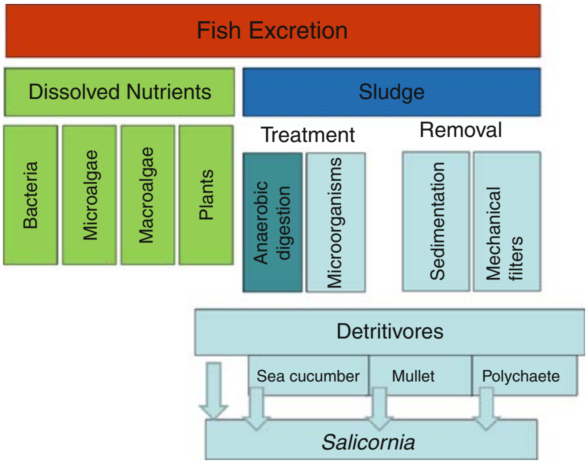
	Assimilation rates	References
Microalgae	1–3 g N m^{-2} day^{-1}	Shpigel and Blaylock (1991)
		Shpigel et al. (1993a)
		Shpigel et al. (2007)
Macroalgae/ *Salicornia*	3–5 g N m^{-2} day^{-1}	Neori et al. (1991)
		Boarder and Shpigel (2001)
		Schuenhoff et al. (2003)
		Neori et al. (2004)
Bivalves/ Artemia	0.3 g N kg^{-1} day^{-1} 6 g N kg^{-1} m^{-3} day^{-1} (20 kg m^{-3})	Shpigel and Blaylock (1992)
		Shpigel et al. (1993a,b, 1994, 1996)
		Zmora and Shpigel (2006)
		Neori et al. (2004, 2006)
Abalone/sea urchins	0.5 g N kg WW day^{-1}	Shpigel et al. (1996, 1999, 2005, 2006)
		Neori et al. (2001)
		Stuart and Shpigel (2009)
Salicornia wetland	2–5 g N m^{-2} day^{-1}	Envirophyte (2010)
		Stuart and Shpigel (2009)

For example, in a fish-bivalve-seaweed IMTA system in Eilat, 63% of the nitrogen from the feed was assimilated by edible organisms, 32% sank to the bottom as biodeposit (sludge), and only 4.1% was discharged back to the sea (Fig. 4) [3].

Nitrogen, phosphate, and silicate ratios can vary according to local farm conditions.

Nutrient composition is also affected by additional biochemical processes in the effluent water such as nitrification, denitrification, and ammonification which occurs in the sedimentation pond as well in the pond walls and in the water pipes. These processes can be accelerated or affected by water temperature, nutrient loads, flow rates, and fish feed biochemical composition. Local natural microfauna in the ponds (e.g., zooplankton) and microflora, as well as bloom and crash phenomena, can affect the water quality as well. In most cases, effluent water from fishponds is characterized by a mixture of ammonia, nitrate, and nitrite.

While macro- and microalgae have proven effective components in land-based systems, neither removes 100% of the dissolved matter and they do not remove particulate matter at all. The remaining waste that includes, among other components, feces, uneaten feed, algae and bacteria, sinks to the bottom and becomes what is known as sludge. This sludge contains valuable ingredients, but can also be toxic to the cultured organisms. It can increase stress and disease risk, and reduce the quality of the water both in situ and for reuse. Ignoring the negative effects of the sludge can thus create serious problems and cause financial losses to the farmers. Removing and dumping sludge into the environment would similarly cause damage, even if moderated by dilution, and "foul the fish farmer's own nest" should he use seawater pumped in from the same area. Using detritivores is a novel option for land-based IMTA. Detritivore organisms such as mullets, cockles, and sea cucumbers will assimilate the waste into their bodies, thereby generating a significant saving in treatment costs, while additionally serving as valuable products in their own right, without requiring the purchase of feed for their culture.



Mariculture Systems, Integrated Land-Based. Figure 4 Different pathways to treat sludge from fishponds

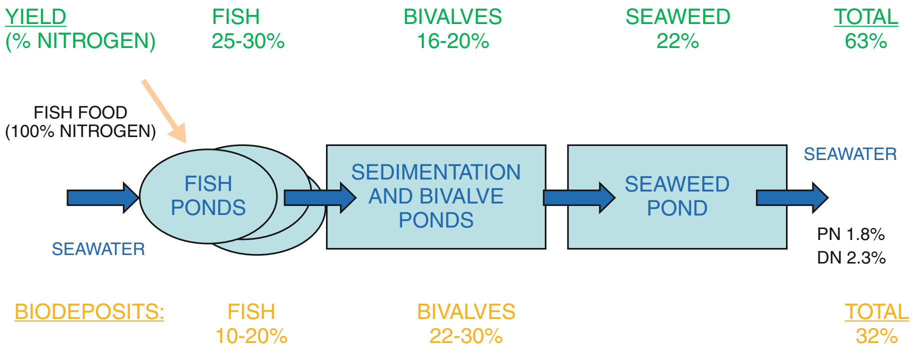
The halophyte *Salicornia* sp. as a biofilter in constructed wetlands was evaluated in the "Genesis" and "Envirophyte" EU projects [34, 45, 46]. Using constructed wetlands (CW) planted with halophytes, which would take up the nutrient-rich wastewater and convert it into valuable plant biomass, is a new option for land-based IMTA. This system was developed to a practical stage for cold (UK) and warm (Israel) water. It was found that CW is efficient in clearing water of nutrients and suspended solids, some materials being purified through incorporation into the plants' biomass and others attaching to the substrate or being broken down by bacteria living therein. CW has the benefit of being low cost, is simple to operate, and can be given an aesthetically pleasing appearance. These plants have commercial value as a health food and are potential candidates for the health, beauty, and nutraceutical industries.

Pilot Scale Systems

In R&D projects in Eilat, Israel, three different types of IMTA systems were developed:

1. Fish (seabream *Sparus aurata*) – seaweed (*Ulva lactuca*)
2. Fish (seabream *Sparus aurata*) – abalone (*Haliotis discus hannai*)/sea urchin (*Paracentrotus lividus*) (macroalgivores) – seaweed (*Ulva lactuca*)
3. Fish (seabream *Sparus aurata*) – bivalve (*Crassostrea gigas* and *Tapes philippinarum*) – seaweed (*Ulva lactuca*)

In the seabream-*Ulva* system, a daily ration of 1.3 t of feed supported 250 t of fish. This amount of food is equivalent to 64 kg of nitrogen. The fish assimilate



Mariculture Systems, Integrated Land-Based. Figure 5
Nitrogen budget of fish-bivalve-seaweed IMTA system in Eilat, Israel

Mariculture Systems, Integrated Land-Based. Table 2 Expected performance of land-based IMTA (WW = wet weight)

IMTA system	Organism	Pond size Ratio/ha	Yield (WW t year^{-1})	Yield (kg WW m^{-2} year^{-1})
Fish-*Ulva* (500 t feed y^{-1})	Seabream	1	220	22
	Ulva	2.5	1,600	64
	Total		1,820	86
Fish-*Ulva* abalone/sea urchin (500 t feed y^{-1})	Seabream	1	220	22
	Ulva	2.5	1,600	64
	Abalone	1.8	185	10
	Sea urchins	1.8	140	8
	Total		1960–2005	94
Fish-*Ulva*-clam/oyster (500 t feed y^{-1})	Seabream	1	220	20
	Clams/*oysters*	4	140	8
	Ulva	0.5	70	64
	Total		430	92

around 16 kg of nitrogen. About 9.6 kg of the nitrogen is drained as particulate nitrogen, and 38.4 kg is drained as dissolved nitrogen. One hectare (ha) of macroalgae (*Ulva lactuca*) is required to remove most of the dissolved nitrogen from the water. This system using 500 t of food per year would require an area of 3.4 ha, at a ratio of 1 ha fish to 2.5 ha *Ulva*. Expected yield is approximately 220 t of fish and 1,600 t of *Ulva* (modified from [5] and [47]) (Table 2).

In the seabream-*Ulva*-macroalgivores (sea urchins/abalone) IMTA system, 1 ha of macroalgae produces 1,600 t of *Ulva* annually. This *Ulva* supports 133 t (WW) of abalone (*Haliotis discus hannai*) or 200 t of sea urchins (*Paracentrotus lividus*). A seabream-*Ulva*-sea urchins/abalone IMTA system in Eilat, Israel, using 500 t of food per year will need an area of 5.3 ha, at a ratio of 1 ha for fish, 2.5 ha for *Ulva*, and 1.8 ha for the macroalgivores (modified from [5] and [47]) (Table 2).

In the seabream, microalgae, and bivalves (*Crassostrea gigas* and *Tapes philippinarum*) IMTA system, a daily ration of 1.3 t of feed supports 250 t of fish. The fish assimilate around 16 kg of nitrogen; 38.4 kg of nitrogen is drained as dissolved nitrogen. This system using 500 t of food per year would need an area of 2 ha of phytoplankton pond (with assimilation efficiency of

1–2 g N m^{-2} day^{-1}) to support production of 140 t bivalves and 70 t of seaweed (modified from [5] and [47]) (Table 2).

The economics of these types of land-based IMTA systems were summarized in [5]. However, the economics of a land-based IMTA are site specific since they depend on variables including local construction and operating costs and market prices for the farm's products at any given time [34].

Additional anticipated parameters based on the same model of using 500 t feed per year in each of the three IMTA systems tested in Eilat, Israel, with the projected yields as depicted in Table 2, can be seen in Table 3.

Mariculture Systems, Integrated Land-Based. Table 3 Anticipated parameters for organisms in the three IMTA systems tested in Eilat, Israel, based on 500 t feed per year

Seabream
FCR = 1.9; Feed protein content = 49%
Fish stocking density = 200 t ha^{-1}; Annual fish yield −300 t ha^{-1}
Seabream farm gate price = €4 kg^{-1}
Seaweed
Ammonia uptake rate –4 g m^{-2} day^{-1}; ammonia uptake efficiency = 85%
Annual Ulva yield = 900 t ha^{-1}
Seaweed (WW) price = €0.5 kg^{-1}
Abalone
FCR = 12; stocking density = 25 kg m^{-2}
Annual yield = 10 kg m^{-2};
Farm gate price = €35 kg^{-1}
Sea urchins
FCR = 8 t *Ulva* 1 t of production; stocking density = 10 kg m^{-2}
Annual yield = 8 kg m^{-2}
Farm gate price = €10 kg^{-1}
Clams/Oysters
Clam annual yield = 6–8 kg m^{-2}
Clams farm gate price = €4.5 kg^{-1}
Oyster annual yield = 25 kg m^{-3}
Oyster farm gate price = €3.5 kg^{-1}

Future Directions: Challenges and Constraints

Although considerable information is already available for putting land-based IMTA systems into practice, much of it is designed around commercial exploitation of a few high value species that are not affordable for the masses. The challenge for the future is to produce a large quantity of aquaculture products that will be cost-effective for producers, at a reasonable price for consumers, and ecologically sustainable.

Additional studies are required to overcome further constraints, including biological, engineering, and economical aspects:

Biological Aspects

- To acquire the knowledge necessary to maintain the correct balance between nutrient production by the system's core organism, nutrient uptake capacity of microalgae and macroalgae, shellfish filtering efficiency, and macroalgivores' activity in the system
- To acquire the knowledge necessary to maintain steady populations of microalgae (mainly diatoms) for the filter feeders and of macroalgae for the macroalgivores within the system in order to avoid blooms and crashes
- To acquire the knowledge necessary for the efficient regeneration of the biodeposit (sludge) from the bottom back to dissolved nutrients for the macro- and microalgae
- To effectively control diseases of the cultured organisms in IMTA systems and transmission of pathogens between components of the system

Engineering Aspects

- To reduce construction and operating costs by engineering improvements
- To minimize heat loss or gain in downstream components of the system
- To increase the use of greenhouse-covered modular systems, gravitation, low head upwelling, water semi-recirculation and other promising energy-saving methods

Economical Aspects

- To render cost effective the use of the extensive areas required for cultivating micro- and macroalgae which cannot be done in a fully recirculating system and for which the facilities must thus be located not too far from the sea

- To develop and diversify the market of seaweed for human consumption from IMTA in Europe and North America
- To develop new markets and consumer acceptance of IMTA products

With the dramatic increase in population and food requirements, traditional extensive production systems cannot satisfy present and future market needs. Modern intensive monoculture systems are not ideal for mass production because they focus on few and expensive species, require high levels of resources, and produce undesirable wastes. To achieve high production rates and environmental conservation, food production using land-based IMTA systems is one of the most promising routes. The IMTA method assimilates expensive nitrogen waste into a valuable product that will increase profit for the farmer, improve FCR, diversify the mariculture products, create additional jobs, and, most importantly, reduce environmental pollution.

Bibliography

1. Goldman JC, Tenore RK, Ryther HJ, Corwin N (1974) Inorganic nitrogen removal in a combined tertiary treatment-marine aquaculture system. I. Removal efficiencies. Water Res 8:45–54
2. Ryther JH, Goldman JC, Gifford JE, Huguenin JE, Wing AS, Clarner JP, Williams LD, Lapointe BE (1975) Physical models of integrated waste recycling-marine polyculture systems. Aquaculture 5:163–177
3. Shpigel M, Neori A, Popper DM, Gordin H (1993) A proposed model for 'environmentally clean' land-based culture of fish, bivalves and seaweeds. Aquaculture 117:115–128
4. Shpigel M (2005) The use bivalves as biofilters and valuable product in land based aquaculture systems-review. In: Dame R, Olenin S (eds) The comparative roles of suspension-feeders in ecosystems. Kluwer, Dordrecht, The Netherlands, p 400
5. Shpigel M, Neori A (1996) The integrated culture of seaweed, abalone, fish and clams in modular intensive land-based systems: I. Proportion of size and projected revenues. Aquacult Eng 155:313–326
6. Troell M, Halling C, Nilsson A, Buschmann AH, Kautsky N, Kautsky L (1997) Integrated marine cultivation of *Gracilaria chilensis* (Gracilariales, Rhodophyta) and salmon cages for reduced environmental impact and increased economic output. Aquaculture 156:45–61
7. Troell M, Norberg J (1998) Modelling output and retention of suspended solids in an integrated salmon-mussel culture. Ecol Model 110:65–77
8. Huguenin JH (1976) An examination of problems and potentials for future large-scale intensive seaweed culture systems. Aquaculture 9:313–342
9. Tenore KR (1976) Food chain dynamics of abalone in a polyculture system. Aquaculture 8:23–27
10. Hughes-Games WL (1977) Growing the Japanese oyster (*Crassostrea gigas*) in sub-tropical seawater fishponds. I. Growth rate, survival and quality index. Aquaculture 11:217–229
11. Gordin H, Motzkin F, Hughes-Games A, Porter C (1981) Seawater mariculture pond – an integrated system. Eur Aquacult Spec Publ 6:1–13
12. Krom MD, Erez J, Porter CB, Ellner S (1989) Phytoplankton nutrient uptake dynamics in earthen marine fishponds under winter and summer conditions. Aquaculture 76:237–253
13. Erez J, Krom MD, Neuwirth T (1990) Daily oxygen variations in marine fish ponds, Elat, Israel. Aquaculture 84:289–305
14. Shpigel M, Fridman R (1990) Propagation of the manila clam *Tapes semidecussatus* in the effluent of marine aquaculture ponds in Elat, Israel. Aquaculture 90:113–122
15. Shpigel M, Blaylock RA (1991) The Pacific oyster, *Crassostrea gigas*, as a biological filter for a marine fish aquaculture pond. Aquaculture 92:187–197
16. Shpigel M, Lee J, Soohoo B, Fridman R, Gordin H (1993) The use of effluent water from fish ponds as a food source for the pacific oyster *Crassostrea gigas* Tunberg. Aquac Fish Manage 244:529–543
17. Neori A, Shpigel M (1999) Using algae to treat effluents and feed invertebrates in sustainable integrated mariculture. World Aquac 302:46–51
18. Neori A, Shpigel M, Scharfstein B (2001) Land-based low-pollution integrated mariculture of fish, seaweed and herbivores: principles of development, design, operation and economics. Aquaculture Europe 2001 book of abstracts, European Aquaculture Soc. Special Publ. No. 29, pp 190–191
19. Lefebvre S, Barille L, Clerc M (2000) Pacific oyster (*Crassostrea gigas*) feeding responses to a fish-farm effluent. Aquaculture 187:185–198
20. Jones AB, Dennison WC, Preston NP (2001) Integrated mariculture of shrimp effluent by sedimentation, oyster filtration and macroalgal absorption: a laboratory scale study. Aquaculture 193:155–178
21. Haines KC (1975) Growth of the Carrageenan-producing tropical red seaweed *Hypnea musciformis* in surface water, 870 m deep water, effluent from a clam mariculture system, and in deep water enriched with artificial fertilizers or domestic sewage. In: 10th European Symposium on Marine Biology, 1, pp 207–220
22. Langton RW, Haines KC, Lyon RE (1977) Ammonia nitrogen produced by the bivalve mollusc *Tapes japonica* and its recovery by the red seaweed *Hypnea musciformis* in a tropical mariculture system. Helgoländer wiss Meeresunters 30:217–229
23. Lapointe BE, Ryther HJ (1978) Some aspects of the growth and yield of *Gracilaria tikvahiae* in culture. Aquaculture 15: 185–193

24. DeBusk TA, Blakeslee M, Ryther JH (1986) Studies on the outdoor cultivation of *Ulva lactuca*. L Bot Mar 29:381–386
25. Bird K (1989) Intensive seaweed cultivation. Aquaculure Mag November/December 1989:29–34
26. Vandermeulen H, Gordin H (1990) Ammonium uptake using *Ulva* (Chlorophyta) in intensive fishpond systems: mass culture and treatment of effluent. J Appl Phycol 2:363–374
27. Cohen I, Neori A (1991) *Ulva lactuca* biofilters for marine fishponds effluents. Bot Mar 34:475–482
28. Neori A, Ellne SP, Boyd CE, Krom MD (1993) The integration of seaweed biofilters with intensive fish ponds to improve water quality and recapture nutrients. In: Moshiri GA (ed) Constructed wetlands for water quality improvement. Lewis, Boca Raton, pp 603–607
29. Schuenhoff A, Shpigel M, Lupatsch I, Ashkenazi A, Msuya FE, Neori A (2003) A semi-commercial, integrated system for the culture of fish and seaweed. Aquaculture 2211–4:167–181
30. Shpigel M, Neori A, Marshall A (1996) The suitability of several introduced species of abalone Gastropoda: Haliotidae for land-based culture with pond grown seaweed in Israel. Israeli J Aquaculture/Bamidgeh 484:192–200
31. Butterworth A (2010) Integrated Multi-Trophic Aquaculture systems incorporating abalone and seaweeds. *A report for Nuffield Australia Farming Scholars, Nuffield Australia Project no. 914*
32. Nobre AM, Robertson-Andersson D, Neori A, Sankar K (2010) Ecological-economic assessment of aquaculture options: comparison between abalone monoculture and integrated mult-trophic aquaculture of abalone and seaweeds. Aquaculture 306:116–126
33. Neori A, Chopin T, Troell M, Buschmann AH, Kraemer GP, Halling C, Shpigel M, Yarish C (2004) Integrated aquaculture: rationale, evolution and state of the art emphasizing seaweed biofiltration in modern mariculture. Aquaculture 231:361–391
34. Bunting SW, Shpigel M (2009) Evaluating the economic potential of horizontally integrated land-based marine aquaculture. Aquaculture 294:43–51
35. Subandar A, Petrell RJ, Harrison PJ (1993) *Laminaria* culture for reduction of dissolved inorganic nitrogen in salmon farm effluent. J Appl Phycol 5:455–463
36. Ahn O, Petrell R, Harrison PJ (1998) Ammonium and nitrate uptake by *Laminaria saccharina* and *Nereocystis luetkeana* originating from a salmon sea cage farm. J Appl Phycol 10:333–340
37. Buschmann AH, Troell M, Kautsky N, Kautsky L (1996) Integrated tank cultivation of salmonids and *Gracilaria chilensis* (Gracilariales, Rhodophyta). Hydrobiologia 326(327):75–82
38. Chopin T, Yarish C (1998) Nutrients or not nutrients? That is the question in seaweed aquaculture... and the answer depends on the type and purpose of the aquaculture system. World Aquaculture Mag 29(31–33):60–61
39. Troell M, Rönnbäck P, Halling C, Kautsky N, Buschmann A (1999) Ecological engineering in aquaculture: use of seaweeds for removing nutrients from intense mariculture. J Appl Phycol 11:89–97
40. Ruokolahti C (1988) Effects of fish farming on growth and chlorophyll content of *Cladophora*. Mar Pollut Bull 19:166–169
41. Rönnberg O, Ådjers K, Roukolathi C, Bondestam M (1992) Effects of fish farming on growth epiphytes and nutrient content of *Fucus vesiculosus* L. in the Åland archipelago, northern Baltic Sea. Aquat Bot 42:109–120
42. Chopin T, Yarish C, Wilkes R, Belyea E, Lu S, Mathieson A (1999) Developing *Porphyra*/salmon integrated aquaculture for bioremediation and diversification of the aquaculture industry. J Appl Phycol 11:463–472
43. Folke C, Kautsky N (1989) The role of ecosystems for a sustainable development of aquaculture. Ambio 18:234–243
44. Soto D (2009) integrated mariculture, a global review. *FAO Fisheries and aquaculture technical paper 529*
45. GENESIS (2004) Development of a generic approach to sustainable integrated marine aquaculture for European environments and markets. (European Economic Community; IPS-2000-102)
46. ENVIROPHYTE (2009) Improvement of the cost effectiveness of marine land-based aquaculture facilities through use of Constructed Wetlands with *Salicornia* as an environmentally friendly biofilter and a valuable by-product. (European Economic Community; SME – 37162)
47. Neori A, Ragg NLC, Shpigel M (1998) The integrated culture of seaweed, abalone, fish and clams in intensive land-based systems: II. Performance and nitrogen partitioning within integrated abalone *Haliotis tuberculata* and macroalgae *Ulva lactuca* and *Gracilaria conferta* culture system. *Aquacultural Engineering* 17(4):215–233

Marine and Freshwater Fecal Indicators and Source Identification

Sandra L. McLellan[1], Alexandria B. Boehm[2],
Orin C. Shanks[3]

[1]School of Freshwater Sciences, University of Wisconsin-Milwaukee, Milwaukee, WI, USA

[2]Environmental and Water Studies, Department of Civil and Environmental Engineering, Stanford University, Stanford, CA, USA

[3]National Risk Management Research Laboratory, United States Environmental Protection Agency, Cincinnati, OH, USA

Article Outline

Glossary

Alternative fecal indicators Fecal indicators that have not been fully validated for standard water quality methods, but show potential for increased sensitivity or specificity over current indicators.

Commensal The general meaning of this word is sharing of food and originates from the Latin word *cum mensa*, meaning "sharing a table." In the context of bacteria and host interactions, the bacteria benefit from the host without causing harm.

Enterococci The term "enterococci" is a general reference to members of the genus *Enterococcus*: however, in the context of water quality standards, enterococci often refers to *E. faecalis* and *E. faecium*, which can be enumerated using selective and differential media.

***Escherichia coli* (*E. coli*)** Gram-negative bacteria found in the gastrointestinal tract of almost all warm-blooded animals. These bacteria are easily cultured and can be enumerated using selective and differential media.

Fecal indicator A chemical or biological constituent that is found in fecal matter that can be used to demonstrate the presence of contamination.

Pathogen A microbe or microorganism such as a bacteria, virus, fungi, prion, or protozoan that causes disease in its animal or plant host.

Polymerase chain reaction (PCR) A scientific technique in molecular biology to amplify a single or few gene copies of a nucleic acid fragment across several orders of magnitude. Amplification results in the generation of thousands to millions of copies of a particular nucleic acid sequence.

Quantitative PCR A technique based on PCR which simultaneously amplifies and quantifies a targeted nucleic acid molecule.

Fecal Indicator Definition

Fecal indicators are organisms or chemical constituents found in fecal material or wastewater that can be measured to demonstrate the presence of fecal pollution. Fecal waste from humans and other animals can contaminant surface waters and pose a serious threat to the environment and human health. Fecal pollution serves as a vehicle for disease transmission including pathogenic bacteria, viruses, or protozoa. Fecal waste also carries with it harmless *commensal* organisms that live in the gastrointestinal (GI) tract and are often used as fecal indicators since they are present in high numbers. The type and amount of pathogens found in fecal pollution is dependent on the host source (human, agricultural animal, wildlife) and the prevalence of illness in the host population. Therefore, employing fecal indicators that provide information about human and other animal contributions is critical for estimating the likelihood that pathogens are present and for directing remediation efforts.

Introduction

Fecal indicators when detected demonstrate the presence of fecal pollution. Fecal indicators play an important role in regulation. Governmental agencies charged with the protection of human health use indicators to assess recreational water quality. In much of the developed world, *Escherichia coli* and *enterococci* are the organisms used for this purpose. Their quantitative link to human health risk in recreational epidemiology studies has led to development of water quality criteria to limit their concentrations in the USA and worldwide.

Conventional indicator methods focus on the cultivation of *E. coli* or enterococci cells isolated from an environmental sample. Culture-based methods are inexpensive and do not require extensive laboratory training to implement. However, these methods are time consuming, requiring 18–24 h to process samples. They also have other limitations such as the inability to discriminate between different animal sources and the potential of indicator microorganisms to persist and sometimes proliferate in the environment. As new scientific discoveries provide a broader view of the different microbes or chemicals associated with fecal pollution and specific sources, new indicators are being identified. These indicators are often referred to as alternative indicators since they have not been fully validated for use for standard methods in water quality testing, but show promise to address some of the limitations associated with conventional fecal indicator approaches.

Some alternative indicators are common to all sources of fecal pollution and can be used as general fecal pollution indicators. Others are associated with a particular host or group of animals. Host-associated indicators are useful for fecal source identification approaches, which are aimed at improving estimates of potential health risk due to pathogens, or identifying major pollution sources that should be remediated. Alternative indicators may take the place of conventional indicators as technology advances. Technologies such as real-time *quantitative PCR* (qPCR), flow cytometry, and advanced chemical analyses can detect previously uncultured microbes or chemicals associated with fecal pollution.

Impact of Fecal Pollution on Coastal Waters

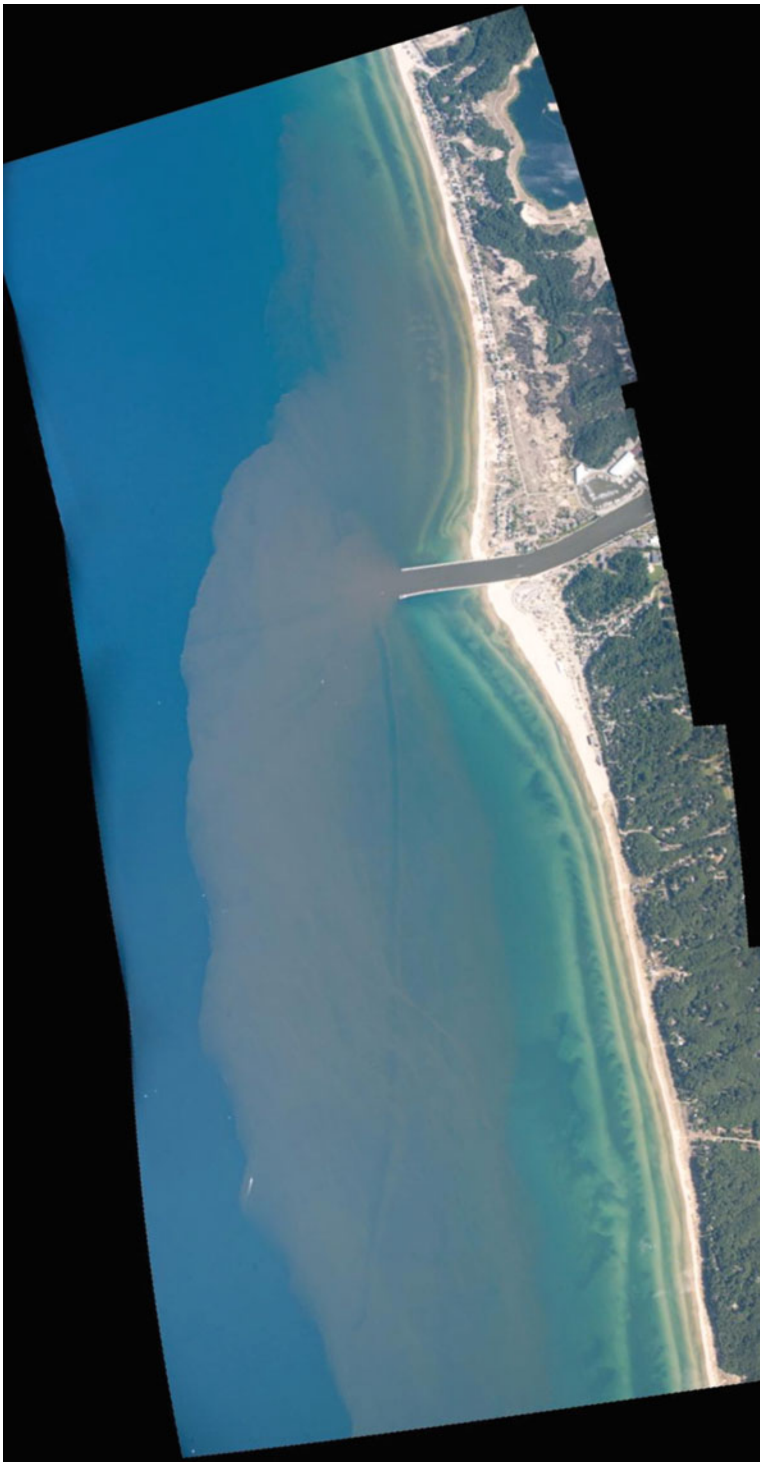
Coastal waters are a valuable resource. Fecal pollution of beaches is not only a threat to human health [13, 52, 98, 195, 258], but can also result in economic losses to surrounding communities [113, 197]. Within the USA, the ocean and Great Lakes coasts encompass more than 15,000 miles of coastline and are the home of economic and recreational centers and unique and rich ecosystems. Many coastal areas are stressed because of dense development and subsequent anthropogenic impacts (Fig. 1). Studies have shown that with increasing urbanization, there is an increase of fecal pollution in waterways [91, 151, 261]. Agricultural land use in upper reaches of watersheds also contributes to fecal pollution in tributaries that ultimately discharge into the ocean or the Great Lakes [24, 261]. Fecal pollution is the major cause of biological water quality impairment in the USA and is the primary cause of recreational beach advisories and closing [252]. Currently, fecal pollution impacts are determined by measuring fecal indicator bacteria using conventional, culture-based approaches. In 2009, there were 18,682 advisories and closures at 2,876 beaches in the USA that are routinely monitored for fecal pollution.

The Link Between Waterborne Disease and Fecal Pollution

Fecal pollution may contain pathogens that can cause disease in humans. To date, there are more than 150 different agents of disease that can be considered waterborne pathogens. This list grows each year as additional emerging pathogens are identified. Table 1 lists common waterborne pathogens and their major host reservoirs. The primary reservoir of human viruses is humans themselves because viruses by nature are host specific; however, animal viruses may also be a concern if they are able to replicate in human hosts. Recent research has identified pigs as a reservoir of hepatitis E virus [99]. Sewage may contain high concentrations of human viruses and some studies have performed surveillance of the viral diseases in the community by monitoring sewage [213, 214]. Some pathogens are predominately found in nonhuman animal hosts, but if humans become infected, person-to-person or waterborne transmission may occur.

Exposure to contaminated water and potential waterborne pathogens most notably causes enteric illness, but skin, ear and eye, or respiratory illnesses may also occur [27, 39, 65, 141, 195, 258]. Many waterborne disease agents are passed through the fecal-oral route, so any activities that involve ingesting contaminated water present a health risk. Ingesting contaminated seafood may also result in exposure to waterborne pathogens (Table 2). For respiratory diseases, inhalation of water droplets or direct contact with mucus membranes can expose a person to a disease-causing agent. Direct contact of contaminated water with wounds could result in an infection.

Recreational waters are of particular concern because swimmers can come into direct contact with contaminated water. Shellfish beds can also be impacted by fecal pollution and are regularly



Marine and Freshwater Fecal Indicators and Source Identification. Figure 1
Plume of river water released into Lake Michigan from the Grand River (Photo provided by Dr. Philip Roberts, Georgia Institute of Technology)

Marine and Freshwater Fecal Indicators and Source Identification. Table 1 Common pathogens responsible for enteric illnesses

Agent		Major sources	References
Viruses[a]			Reviewed by Fong and Lipp [76]
	Enteroviruses	Human	[32, 91, 122, 125, 167, 178]
	hepatitis A	Human	[86, 91, 122]
	Adenoviruses	Human	[51, 108, 122, 124, 125, 271, 273]
	Norovirus	Human	[17, 91, 263, 273]
	Rotavirus	Human	[85, 202]
	Astrovirus	Human	[47, 194]
Bacteria			
	Pathogenic *E. coli*	Humans and animals[b]	[5, 102, 109, 171, 260]
	Shigella	Humans and animals	[118]
	Vibrio cholerae	Humans and environment	[129, 130, 147]
	Campylobacter	Animals and humans	[8, 33, 112, 183, 254, 260, 270]
	Salmonella	Animals and humans	[21, 100, 144, 254]
	Yersinia enterocolitica	Animals	[104, 205]
	Aeromonas	Environment	[192, 215]
	Plesiomonas	Environment	[11, 166, 203]
	Vibrio parahemolyticus	Environment	[55, 127, 254]
Parasites			
	Cryptosporidium parvum	Animals and humans	[90, 112, 144, 149, 270]
	Giardia lamblia	Animals and humans	[90, 112]
	Entameba histolytica	Animals and humans	[152]

[a] Humans are the predominate source of human viruses, but in some cases, transmission from animals to human is possible. For example, this is suspected to be possible for hepatitis E [99], a calicivirus in the same virus family as norovirus

[b] Animals are the primary reservoir of *E. coli* O157:H7 [46], one strain of shiga-toxin-producing *E. coli*

monitored to assure that harvested shellfish has not been subjected to contamination. In the Great Lakes, nearshore coastal waters are a drinking water source to nearly 40 million people. Stringent treatment requirements provide safe drinking water, but both source water and treated drinking water are closely monitored for evidence of fecal pollution to assure that treatment protocols are adequate. For a more complete discussion of this topic, see Chaps. 3–5 of this volume.

Important Attributes of Indicators. Fecal indicators can either be general indicators of fecal pollution or associated with a particular animal source. Many watersheds and coastal waters have mixed land use; therefore, both general fecal indicators and source-associated indicators have an important role in assessing water quality. Ideally, fecal indicators should be present in high levels in fecal pollution so that they can be used as a sensitive measure of the level of contamination when diluted to small concentrations in the environment. Fecal indicators should provide information about host source contribution when possible, whether it is from humans, or different agricultural animals or wildlife. Detection methods should be relatively simple and affordable considering

Marine and Freshwater Fecal Indicators and Source Identification. Table 2 Common pathogens responsible for respiratory illness and skin infections

Agent		Primary source	References
viruses			
	Adenovirus	Humans	[51, 108, 122, 124, 125, 271, 273]
Bacteria			
	Vibrio parahemolyticus	Environment	[55, 127, 254]
	Vibrio vulnificus	Environment	[187, 264]
	Leptospira	Animals (wildlife)	[96, 236]
	Legionella spp.	Environment	[82, 186, 236]
	Staphylococcus aureus	Humans	[48, 49, 88, 254]

For an expanded list, please see [27, 168]

that much of the hands-on monitoring is done by local health departments. Methods should lend themselves to rapid testing so that beach notification can happen in a timely manner.

Clearly, no single indicator can meet all of these goals. Therefore, it is critical to have multiple indicators that can be used in concert if needed. Different indicators will behave differently in various environments, e.g., marine waters versus freshwater [9, 92, 180, 227, 229]. Certain indicators may be appropriate for investigating sources of fecal pollution, or setting remediation goals, whereas others are better suited for rapid detection for recreational water quality monitoring for any fecal pollution present. Water resource managers, public health officials, and researchers must work together to identify what information is needed and choose the most appropriate indicators. For example, *E. coli* is recommended for freshwater, but it has a very short half-life in the open waters of the Great Lakes [160]; therefore, highly persistent indicators such as *Clostridium perfringens* may be more useful for long-term monitoring [71, 146, 169]. Enterococci qPCR is being developed for rapid beach testing [93, 106], but is a general indicator and host-associated indicators within the order *Bacteroidales* may be more useful for identifying sources [31, 72, 236].

Important Attributes of Fecal Indicators
Routine monitoring: Specific for fecal pollution
Present when pathogens are present
Correlate well to illnesses
Easy to quantify
Simple and cost-effective methodology
Behave in the environment in the same manner as pathogens
No growth outside of the host environment
Amenable to rapid detection methodologies
Investigations: Specific for a source of fecal pollution
Sensitive, e.g., present in high numbers in most animals with a given source
Known or predictable ecology
Known or predictable relationship to pathogens

Detection of Conventional Indicators

Common fecal indicators that are used for water quality monitoring or recreational beaches are listed in Table 3. All of these indicators were originally identified as constituents of fecal pollution using selective and differential culture techniques. The earliest methods data back to late 1800s and early 1900s [14, 66] for coliform bacteria. There are two culture approaches for enumerating bacteria in water samples. The most probable number (MPN) methods involve culture-based detection in liquid broth using a series of dilutions. The dilutions in which organisms are detected can be used to calculate a statistical estimate of *enterococci* concentration for that sample. The second approach involves filtering samples through a membrane filter. The filter is transferred to solid selective media that is optimized for the growth of the target organisms and inhibitory for other organisms. Various chromogenic substrates or pH indicators can be incorporated to make the media differential for fecal indicator microorganisms. A review of conventional and novel indicators can be found in Edge and Boehm [69].

Marine and Freshwater Fecal Indicators and Source Identification. Table 3 Historic and conventional fecal indicators

Organisms	Use	Limitations
Total coliforms	Early indicator used in surface waters, currently in use for drinking water[a] since detection of total coliforms provides information on general sanitation	Not specific for fecal pollution
Fecal coliforms	Used for recreational waters until late 1980s or 1990s in some US states, still in use as a standard for wastewater, recreational waters, and shellfish	Some fecal coliforms can grow in the environment
E. coli	Currently recommended by the USEPA for fresh recreational waters[b]. Replaced fecal coliforms as a more specific indicator of fecal pollution	More specific than fecal coliforms, but has been reported to persist and grow in the environment [7, 117, 134, 265]
Fecal streptococci	Early indicator for surface water quality	Not all are fecal specific
Enterococci	Currently recommended indicator for marine recreational waters. Replaced fecal streptococci	General indicator; some grow in the environment
Clostridium perfringens	Proposed in 1963 as an indicator of wastewater and receiving waters. Used in some European countries but not the USA	May survive for long periods in the environment.

[a] Total coliforms (TC) are used since they are a good measure of bacteriological contamination, regardless of fecal or environmental sources. New proposed rules would change the standard from TC to *E. coli*, e.g., more specific for fecal pollution

[b] Criteria are being revised; new criteria will be based on detection of enterococci by culture or qPCR (total *Bacteroides*)

Culture-based methods continue to be widely used for detection of fecal indicator bacteria; however, the time required to obtain a result is a major limitation of these methods for providing rapid (e.g., 4 h) results of beach water quality to assure timely public notification. Molecular methods such as qPCR can be used for detection of traditional fecal indicators [106, 182, 267].

Coliforms. Coliform bacteria are a group of bacteria that were the first indicators of fecal pollution. Coliforms are gram-negative, rod-shaped, facultatively anaerobic, non-spore-forming bacteria found in warm-blooded animals, as well as in soil, water, and vegetation. Coliforms are not a specific taxonomic group of bacteria, but are classified based on a number of characteristics. These organisms are identified by fermentation of lactose with the production of acid and gas at 35–37°C. Coliforms are also negative for cytochrome oxidase and positive for β-galactosidase. Coliforms are measured by using an MPN [16] or by enumeration of colony-forming units (CFU) using membrane filtration and selective and differential media such as MI [249]. These organisms generally are within the family *Enterobacteriaceae* and include the genera *Citrobacter*, *Escherichia*, *Enterobacter*, *Hafnia*, *Klebsiella*, and *Serratia*. Coliform bacteria were one of the earliest indicators of water quality used in the USA, with individual states setting limits of 50–2,400 coliforms per 100 ml of water as a standard for recreation waters in the 1950s and 1960s [66].

Fecal coliforms. Fecal coliforms are a subgroup of coliforms and refer more specifically to coliforms derived from feces. Like coliforms, they are not a specific taxonomic group; they are based upon several morphological and physiological characteristics. These are defined by the same criteria as coliforms, but are thermotolerant and will grow at 44.5°C. *E. coli* is one of the major fecal coliforms found in feces, in addition to members of *Klebsiella*, *Enterobacter*, and *Citrobacter*. The designation of fecal coliforms was intended to improve specificity; however, some organisms included in this group can be found free living in the environment, most notably *Klebsiella* [42, 83, 177]. Beach water samples have also been found that have evidence of fecal coliforms that have replicated in the environment [158].

The first national water quality criterion for recreational waters was based upon fecal coliforms. In 1968, the National Technical Advisory Committee, commissioned by the Federal Water Pollution Control Administration (now referred to as the Environmental Protection Agency), determined that 400 fecal coliforms per 100 ml corresponded to an adverse GI health effect [66]. Subsequent recommendations stated that for recreational waters, within a 30-day period, the geometric mean should not exceed 200 fecal coliforms per 100 ml, and 10% of the samples should not exceed 400 fecal coliforms per 100 ml. Fecal coliforms are no longer used for recreation waters in most states, but the basis of the 1968 criteria is still used for regulating water quality of wastewater treatment plant effluents and for assessing river water quality. Fecal coliforms are also still used for shellfish testing (water overlying the reefs and oyster meats).

Escherichia coli (E. coli). *E. coli* is a fecal coliform that has been suggested to be more specific for fecal pollution than testing for the group of fecal coliforms and was recommended as an indicator for freshwater in 1986 by the United States Environmental Protection Agency (USEPA) [14, 247]. *E. coli* are present in the GI tract of most warm-blooded animals, and therefore a general indicator of fecal pollution. *E. coli* is a thermotolerant coliform that produces indole from tryptophan and it can be differentiated from other microorganisms based on β-glucuronidase activity. Selective and differential media tests for this activity using methods based on membrane filtration, modified mTEC [248], or MPN approaches such as the Colilert manufactured by IDEXX [68] are commonly used to identify *E. coli* in surface water samples. One testing methodology simultaneously detects coliforms and *E. coli* using β-galactosidase and β-glucuronidase activity, respectively, as discriminators [249]. Some epidemiology studies have shown a relationship between *E. coli* densities and GI illness [65, 195]. *E. coli* has some limitations as a fecal indicator at recreational beaches because it has been shown to persist and even grow in some aquatic environments, thereby potentially interfering with the relationship between *E. coli* and recent fecal pollution events [7, 26, 134, 265].

Enterococci. Enterococci are gram-positive cocci and are nearly universally present as commensal organisms in the intestine of human and nonhuman animal hosts. The most common species in human hosts are *E. faecalis* and *E. faecium* [58, 139]. The enterococci are a subgroup of the fecal streptococci. Fecal streptococci have also been referred to as Group D streptococci according to Lancefield serotyping. The fecal streptococci have historically been used as fecal indicators and include species from two genera: *Enterococcus* and *Streptococcus.* There are two *Streptococcus* species in the fecal streptococci group – *Streptococcus bovis* and *Streptococcus equinus* – that have been shown to survive poorly in water. Hence, in water, fecal streptococci and enterococci are thought to be equivalent [116].

In the USA and the EU, enterococci are used for monitoring marine bathing waters because epidemiology studies have linked their concentration to human health outcomes [256]. The standards are tied to approved culture-based methods for their quantification: multiple-tube fermentation, membrane filtration, and defined substrate assays. Clesceri et al. (1998) describe a multiple-tube method where azide dextrose broth is used followed by confirmation with Pfizer selective *Enterococcus* (PSE) media and brain-heart infusion broth with 6.5% NaCl. Both the EU and the USA have approved the use of defined substrate assays manufactured by IDEXX for the quantification of enterococci (Enterolert and Enterolert-E). The USEPA-approved method 1600 utilizes membrane filtration onto mEI media for quantification [250]. Studies that have compared these culture-based methods for quantifying enterococci often find the methods yield slightly different results [32].

The USEPA has developed a qPCR assay for the enumeration of enterococci which has been compared to membrane filtration results [106, 253]. Enterococci measured via qPCR often yield higher concentrations than culture-based measurements since it enumerates both live and dead bacteria [32]. Enterococci measured by qPCR have been linked to human health outcomes in epidemiology studies of marine and fresh water beaches [257–259]. Ongoing work is focused on better defining these links. As the USEPA formulated new recreational water quality criteria, qPCR for enterococci is expected to be included as a rapid method which allows beach managers and public health workers to post water quality advisories on the same day the sample is taken.

Clostridium perfringens. *C. perfringens* is member of the phylum *Firmicutes* and is a gram-positive, low GC content organism. *C. perfringens* was suggested as a potential indicator in 1963 [34], and gained acceptance in EU countries, but it was not chosen for use in the USA because it survives for long periods of time in the environment [14, 66]. Epidemiology studies report a relationship between *C. perfringens* and illness [268], while other studies found no relationship [39]. However, *C. perfringens* has been shown to be a useful fecal indicator in certain environments where other indicators are highly modulated by environmental factors. Studies in tropical waters suggested *C. perfringens* is a better indicator compared with fecal coliforms because it is a spore-forming organism and does not replicate in the environment [79, 80]. Because of its spore-forming ability, *C. perfringens* has been used as a tracer of long-term fecal pollution impacts in marine and freshwater systems [36, 54, 70, 110, 169]. *C. perfringens* has also been suggested as a good indicator in the open waters of the Great Lakes because it can serve as a conservative tracer of fecal pollution and may mimic protozoan cyst or oocyst survival [164, 169, 191].

Alternative indicators. Ongoing research studies have identified a broad array of new potential indicators of fecal pollution. Molecular-based methods have made possible the characterization of organisms that previously were either not recognized as associated with fecal pollution, or were difficult to detect due to complex cultivation requirements. Alternative indicators may also employ unique chemical constituents. Alternative indicators are being developed as general detection of fecal pollution, such as total *Bacteroides* [53, 59], as well as source identifiers associated with a particular animal group (Table 4).

Different sources of fecal pollution can contribute different types and concentrations of pathogens (Table 1 and Table 2). For example, human fecal sources, particularly sewage, contain waste from a large number of people and are considered a primary source of human enteric viruses. *Cryptosporidium* may be associated with cattle waste. Fecal indicators that provide information about the source will improve our ability to estimate the health risk due to pathogens as well as direct remediation efforts to major contributing sources of fecal pollution. The development of qPCR methodology has also advanced simple presence/absence detection to quantitative estimates of fecal pollution and provides a platform for the implementation of rapid methods.

16S rRNA gene targets. Many of the alternative indicators that have been described are based on detection of the organisms based on the 16S rRNA gene sequence. This gene is highly conserved among bacteria and has been used extensively to assign taxonomy.

Bacteroidales. Members of the order *Bacteroidales* are potentially useful indicators of fecal contamination because they generally are found in high numbers in fecal material of humans and other warm-blooded animals and are unlikely to survive in the beach environment [74, 136]. Early studies identified unique sequences in the *Bacteroides* 16S rRNA gene from human and ruminant *Bacteroides* species that are associated with respective fecal pollution sources [24, 136]. Sequencing of clone libraries demonstrated that sequences of members of the broader *Bacteroidales* group, rather than exclusively *Bacteroides* spp., are amplified with primers originally targeting total *Bacteroides* spp. [60]. Subsequent studies have used taxon-specific cloning to characterize *Bacteroidales* populations within humans and different animals and have identified a broad range of host-associated genetic markers [25, 60, 75, 78, 121, 133, 137, 138, 153, 165, 184]. Since culture techniques for isolation of these anaerobic bacteria are difficult to perform, molecular techniques have been developed to amplify, detect, and in some cases quantify the 16S rRNA genes of *Bacteroides* spp. from feces and ambient water [53, 59, 133, 153, 218, 262]. Many of these assays utilize the HF183 sequence first reported by Bernhard and Field [25]. The utility of the genetic markers has been tested extensively in fecal impacted environments, including beaches [1, 37, 84, 181, 207, 216]. In addition, numerous studies report information on the distribution of these host-associated genetic markers in target and non-target populations [3, 64, 133, 138, 143, 185, 224, 225, 228], relationship to pathogens [208, 209], and the decay of these genetic markers in marine and freshwaters [20, 61, 184, 210, 261].

Bifidobacterium. This genus represents another group of GI bacteria with particular species reported

Marine and Freshwater Fecal Indicators and Source Identification. Table 4 Examples of biological *alternative fecal indicators* that provide animal host information

	(Notes)	References
Organisms – 16S rRNA gene		
Bacteroidales	*Bacteroidales* associated with hosts have been identified	[25, 60, 75, 78, 133, 137, 138, 153, 165, 184]
Bifidobacterium		[35, 64, 148, 173]
Methanobrevibacterium	Member of Archaea and dominant in the GI	[128, 243–245]
Faecalibacterium		[161, 279]
Lachnospiraceae		[161, 175]
Ruminococcace		[161]
Functional genes		
esp gene in enterococci	Gene responsible for attachment on human epithelial cells	[93, 212]
Toxin genes in *E. coli*		[50, 131]
Beta-glucuronidase	Polymorphisms in this gene have been linked to different host types	[200]
Unknown genes/regions		
Cattle and human markers	Identified by genomic fragment enrichment	[220, 221]
Gull marker	Identified by subtractive hybridization	[101]
Phenotypes		
Antibiotic resistance of standard fecal indicators	Based on the theory that host exposed to antibiotics will have a higher percentage of antibiotic-resistant *E. coli* or enterococci	[97, 105, 188, 269]
Viruses		
F+ coliphage	Type I and IV associated with animals and II and III associated with humans	[114]
Bacteroides phage		[12, 126, 240]
Adenovirus, enterovirus, polyomaviruses	Viruses are host specific by nature, and therefore, detection of human viruses demonstrates human sources are present.	[6, 123, 162, 179, 193]

to be associated with human fecal pollution including *B. adolescentis*, *B. dentium*, and *B. longum* [35, 148, 155, 173]. Several technologies targeting *Bifidobacterium* genes are reported for multiplex PCR detection [35] and qPCR [150, 154]. *Bifidobacterium* typically occur at lower concentrations than *Bacteroidales* making them harder to detect in dilute ambient water samples [219] and exhibit a rapid decay based on bench-scale survival studies [201]. Thus, the detection of a *Bifidobacterium* host-associated genetic marker in a polluted water sample suggests a recent, high concentration contamination event.

Faecalibacterium. This genus of bacteria has been reported in humans and other animals and has been suggested as a potential target for development of host-associated genetic markers [81, 161, 246, 279]. Sewage and cattle have been shown to have a high abundance of *Faecalibacterium* [161, 226]. Additional characterization of this group is needed to characterize phylotypes that are associated with specific animal sources.

Lachnospiraceae. *Lachnospiraceae* are found in high abundance in human fecal samples [57, 77, 242], sewage [161], and cattle [226]. *Lachnospiraceae* are included in the group *Clostridium coccoides* [107, 150].

The proportions of *Lachnospiraceae*, *Bacteroides*, and *Bifidobacterium* of the human microbiota vary among different animal species, and quantification of these proportions has been proposed as a method for fecal pollution source identification [81]. Additional characterizations of this group are needed to characterize phylotypes that are associated with specific animal sources [161].

Gene product targets. Molecular methods have also allowed for detection of genes that serve a functional role in the organism. In some cases, the function may be linked to specific host microbe interactions, making these genetic markers potentially good host-associated alternative indicators [221]. Genetic markers have been identified with a variety of molecular methods, including subtractive hybridization, genome fragment enrichment, and other metagenomic approaches.

Toxin genes *of* E. coli. Specific subpopulations of *E. coli* contain genes coding for toxins, including heat-labile enterotoxin (LT) and heat-stable enterotoxin (ST). *E. coli* carrying toxins are generally clonal populations that are found within certain animal reservoirs and have been suggested as host-associated indicators. Specific sequences of the STII toxin gene were found to be associated with swine, but not present in sewage or dairy farm lagoons [132]. Cattle-associated LTIIa has also been reported [50, 131]. These toxin genes have a worldwide distribution [72]. The occurrence of *E. coli* positive for STII or LTIIa can be low in agricultural animal populations, potentially limiting the use of these genes for the identification of specific animal sources.

Esp *gene.* The enterococcal surface protein (*esp*) gene is a putative virulence factor in *Enterococcus faecium* that has been shown to be associated with enterococci from human origin [212]. Because this gene occurs at a low frequency, original detection methods involved an enrichment step where DNA is extracted from enterococci grown on selective media, followed by PCR. Comparison studies have shown the *esp* gene in enterococci to correlate with other human-associated genetic markers [4, 163] and this alternative indicator has been employed in numerous field studies [275]. Newer methods employ qPCR that can directly detect the *esp* gene [2].

gyrB. The genetic locus *gyr*B is a housekeeping gene (e.g., common to all bacteria because of a central function). Similar to 16S rRNA gene loci, housekeeping genes are generally highly conserved and therefore useful for identifying specific phylotypes. One study employed qPCR targeting *gyr*B in *Bacteroides fragilis* as an indicator of human specific fecal contamination [142].

Methanogens. Methanobrevibacter smithii is a dominant Archaea in the human gut [67]. The *nif*H gene of this organism has been used as a human-associated indicator [243]. Similar assays employing the same gene in *Methanobrevibacter ruminantium* have been developed [245]. Assays for quantification of the *nif*H target have also been developed [22, 128]. An Archaea genetic marker may prove useful because it may have a different survival or ecology compared with bacterial indicators and pathogens.

Bacteroides thetaiotaomicron. *B. thetaiotaomicron* is found in high numbers in humans compared with other animals and is described as a niche organism in the human gut [274]. A genomic fragment that was generated with universal primers as a second unexpected amplicon was found to distinguish *B. thetaiotaomicron* from other animal species [241]. PCR primers specific for *B. thetaiotaomicron* were developed based on the sequence of this 547-bp genomic fragment and have been tested against a number of fecal samples from humans and nonhuman sources [45, 241]. A putative gene for complex polysaccharide degradation has also been used as a genetic marker for qPCR since the trait is hypothesized to be involved in host-associated metabolic pathway [277].

Metagenomics. The majority of host-associated genetic markers available to date target the 16S rRNA gene from a limited number of different microorganisms. Advancements in DNA sequencing and sorting technologies now allow researchers to survey the entire genome of all members of fecal microbial community. Different strategies include the use of competitive hybridization approaches [101], microarrays [145, 272], and 454 pyrosequencing [161, 226, 246]. Whole genome and community approaches vastly expand the number of candidate source-associated genetic markers and may allow for the development of even more refined source identification methods.

Viruses

F specific (F^+) coliphages. F^+ coliphage RNA coliphages have serologically distinct groups that predominate in humans (groups II and III) which are distinct from those commonly found in other animals (group I and IV) [114]. Comparison studies of different alternative indicators suggest F^+ coliphage types are reliable indicators of host sources, but the groups are not exclusive to either animal or human sources [28, 180]. Further, differential survival may influence source identification in natural waters and may need to be taken into account in interpreting source identification studies [38, 172]. However, viral indicators may correlate more closely to human viral pathogens as they may have a similar ecology in the environment.

Bacteroides *phages*. Phages infecting *B. fragilis* and *B. thetaiotaomicron* have been used as indicators of human fecal pollution [12, 126, 240]. Differential ability of host strains of *Bacteroides* to detect phages from different sources has been reported [196], as well as geographic variability. Culture methods have been developed to isolate diverse host *Bacteroides* strains [190]. In survival studies, two *B. fragilis* phages were shown to survive longer in seawater compared to MS2 coliphage [157].

Human polyomaviruses. Human polyomaviruses are widespread among human populations and have been suggested as indicators of human waste [6, 162]. This virus is excreted in the urine and therefore may be detected in the absence of human feces. Studies have compared detection of human polyomaviruses with detection of human *Bacteroides* HF183 genetic marker and enterococci carrying the *esp* gene and found a strong correlation [4, 163].

Chemicals. Chemical methods do not detect fecal bacteria. Instead, these methods are designed to detect chemical compounds associated with human activities or sanitary sewage. Chemical indicators may provide additional evidence as to source [87, 95]. These chemicals are often found in sewage treatment facility discharges and septic tank effluent. For example, optical brighteners are commonly found in laundry detergents and have been used to indicate the presence of human fecal pollution in environmental waters [41, 62]. Fecal sterols such as coprostanol are also reported to be associated with human fecal pollution [28, 115, 176, 239]. Other potential chemical fecal indicators include antibacterial compounds, pharmaceuticals, and caffeine [95, 278] (Table 5).

Quantification of Bacterial Indicators Using qPCR. Conventional or endpoint PCR allows for the selective amplification of a particular genetic marker at extremely low concentrations even in the presence of a mixture of heterologous DNA targets making it ideal for environmental applications. The final result of an endpoint PCR method is either the presence or absence of the DNA target. Even though the qualitative determination of fecal pollution in a water sample can be very useful information, researchers quickly recognized the added advantage of generating quantitative data. The ability to estimate the concentration of a DNA target in a known volume of water provides a means to investigate relationships between the concentration of a fecal indicator genetic marker and numerous factors such as illness rates in swimmers or efficiency of waste management practices.

qPCR relies on the continuous monitoring of PCR product accumulation as amplification occurs. Estimation of the concentration of a genetic marker is based on the theoretical premise that there is a log-linear relationship between the starting amount of DNA target in a reaction and the fractional thermal cycle where PCR product accumulation is first significantly detectable (Table 2); for review see [204]. qPCR applications designed to estimate fecal bacteria concentrations in recreational waters are gaining widespread attention due to the rapid nature of these methodologies (same day results), reports linking the occurrence of DNA targets to public health risk [106, 257, 258],

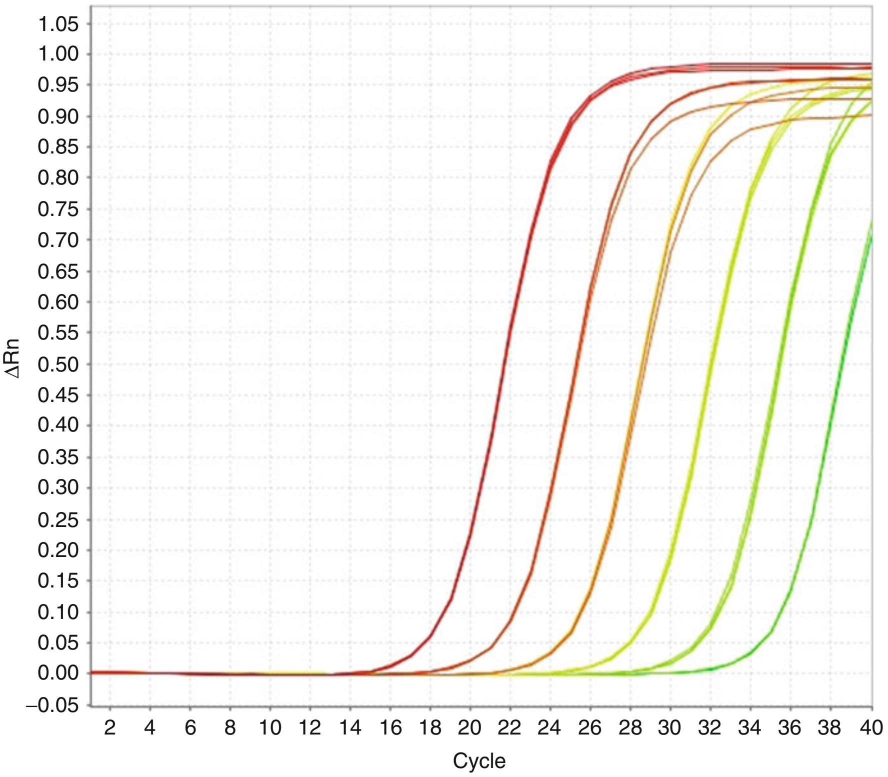
Marine and Freshwater Fecal Indicators and Source Identification. Table 5 Chemical alternative fecal indicators

Chemical constituents
Fecal sterols
Optical brighteners
Caffeine
Personal care products and pharmaceuticals

and the development of host-associated fecal source identification assays [40, 133, 135, 138, 163, 185, 218, 222, 223]. However, there are many technical concerns that must be addressed before these qPCR applications are ready for implementation.

It is important to recognize that a qPCR method consists of several protocols linked in succession including sample collection, sample preparation, nucleic acid purification, target amplification, and data interpretation. Each of these steps plays a critical role in the successful estimation of a DNA target concentration in an environmental sample. In addition, the extremely high level of sensitivity make qPCR methods highly susceptible to cross-contamination during field sampling, nucleic acid purification, and genetic marker amplification (Fig. 2). As a result, numerous studies have been conducted to address issues such as density and distribution of genetic markers in primary and secondary sources [60, 133, 199, 224, 225, 228], sample matrix interference during qPCR amplification [140, 198, 224, 255], estimating decay rates of DNA targets in ambient water [18, 23, 184, 261], loss of target DNA during nucleic acid recovery [106, 170, 238], and selection of a mathematical model to transform raw qPCR data into an estimation of concentration [230, 231].

Microbial Source Identification. Identification of the sources of fecal pollution is important for both developing remediation strategies and for estimating the likelihood of pathogen occurrence. In most cases, the source of fecal pollution in a water body of interest is originally measured because of high amounts of conventional general fecal indicators (i.e., enterococci or *E. coli*). Methods and study designs for source identification, also referred to as “microbial source tracking” (MST) or “fecal source identification” (FSI), has been reviewed extensively [72, 206, 237].



Marine and Freshwater Fecal Indicators and Source Identification. Figure 2
Quantification of real-time polymerase chain reaction (qPCR) product can be achieved by observing an increase in fluorescence, indicating product formation, in relation to cycle number

Identifying fecal pollution sources involves understanding both the physical location of the inputs and the contributing host sources. Most source identification studies begin with spatial and temporal sampling since fecal pollution sources are rarely constant and the locations of inputs are not always obvious. Following release into the environment, the ecology of fecal indicators is greatly influenced by the residence time, type of water body (e.g., marine or freshwater, oligotrophic, or nutrient rich), predation, or even potential growth by some conventional indicators [31, 236]. Therefore, it is very difficult to take one or two samples and determine the major source contributing fecal pollution to an impacted body of water.

Spatial and temporal surveys are complemented by using alternative indicators that can provide information as to the host source of fecal pollution. Often, a first tier assessment will involve distinguishing human versus nonhuman fecal pollution [89, 181]. Cross reactivity needs to be considered, along with geographic relevance of a particular indicator. The possible fecal pollution sources within the watershed need to be considered when choosing the most appropriate alternative indicators. The use of alternative indicators for microbial source identification has been reviewed extensively [20, 72, 73, 206, 211, 235, 237].

Early approaches to microbial source identification focused on library-based methods, where either phenotypic traits or genotypes of indicator bacteria were characterized from a particular source and then compared to what was found in surface waters. Methods for characterizing *E. coli* or enterococci have included antibiotic resistance, ribotyping, and repetitive extragenic palindromic PCR [43, 44, 63, 103, 105, 159, 189, 217, 269]. There are multiple complications in using library-based methods that include applicability of the library across geographic locations, specificity of *E. coli* or enterococci indicators to a particular animal host, and complex genetic relationships among these indicators [10, 72, 159, 206, 236]. Further, creating a library is expensive and multiple water samples need to be analyzed because fecal pollution inputs are usually driven by storm events and can involve multiple animal sources. Most source identification methods have moved to marker-based, or non-library dependent, approaches. Marker-based approaches involve utilizing a chemical or biological constituent that is commonly found in the fecal pollution source of interest, in high abundance so that it can be detected easily and associated with a specific human or animal source (Figs. 3 and 4).

Ecology of Pathogens and Indicators in the Environment

The identification of a host-associated marker of fecal pollution goes beyond microbiology. Once the fecal indicator is discharged into the environment, it becomes necessary to understand the various fate and transport mechanisms that control the concentrations of indicators and pathogens at the point of sampling.

Fate processes include dark or photo-inactivation [32], growth [111], sorption and desorption to sediments [19, 94], and grazing by zooplankton [30]. Inactivation has received the most attention of these fate processes. Although a fair amount of work has examined the interaction of pathogens and indicators with sediments, the work has primarily been focused on porous media, and simplified conditions. More work on the interactions of microbial pollutants and sediments and particles in surface waters is needed, particularly given the widespread occurrence of some indicators and pathogens in sediments and beach sands [7, 26, 56, 117, 265, 275].

Transport processes that control indicator and pathogen transport in surface waters include advection and dispersion of waterborne organisms. These processes are fairly well understood [174] and once determined in a particular surface water, they can be used to model microbial pollution. The resuspension and deposition of sediment-bound organisms is more complicated. Some work has examined these processes for *E. coli* [119, 120] and fecal coliforms [234] in streams and lagoons. Yamahara et al. [275] present a conceptual model for how enterococci in beach sands are suspended into the water column. A better mechanistic understanding of how organisms in the sediment or sand are transported into the water column is warranted.

Of the fate and transport processes described above, perhaps the most important to consider when choosing an indicator for microbial source identification is the time scale of inactivation and its tendency to sorb to sediments. For example, if the goal is enterococci

Marine and Freshwater Fecal Indicators and Source Identification. Figure 3
Stormwater outfalls introduce fecal pollution from domestic pets and wildlife into rivers. Stormwater systems can also become contaminated with human sewage from leaking sanitary sewer systems (Photos provided by Dr. Sandra McLellan, University of Wisconsin-Milwaukee)

Marine and Freshwater Fecal Indicators and Source Identification. Figure 4
Large gull populations are common non-point sources of fecal pollution on beaches (Photos provided by Dr. Sandra McLellan, University of Wisconsin-Milwaukee)

source identification for designing remediation strategies, then ideally, the persistence of the genetic marker will mirror that of enterococci. A health-protective goal may be to have no feces present in a water body. If this is the case, then a source identifier with very long-persistence may be needed. A source identifier that interacts strongly with sediments may be problematic as it may allow sediments to become a secondary, environmental source of the marker. Generally, sediments are believed to be a protective environment for microorganisms, particularly bacteria, where they may persist or even grow [276]. Future work on source identifiers will need to document the importance of sorption and interactions with sediments in general.

Estimating Risk of Pathogen Exposure Using Fecal Indicators

Using fecal indicators to link the presence of fecal pollution to waterborne disease risk is challenging. The types of pathogens that might be present will depend primarily on the source of fecal pollution. For example, sanitary sewer discharges (human sources) may contain high levels of human viruses, whereas wildlife is less likely to carry human viruses, but may contain protozoan and bacteria that can infect humans. Comprehensive models that integrate data from several research fields such as occurrence of pathogens in fecal sources, dose–response relationships, source identifier decay behaviors, acceptable health risk, and route of transmission can be used to estimate risk and are termed quantitative microbial risk assessment (QMRA) [15, 232, 233]. The type of pathogen present will also depend on the prevalence of the disease-causing agent within the population at the time of contamination. Many human viruses are seasonal, and protozoans such as *Cryptosporidium* are prevalent during certain times of the year, such as spring when calves can shed high concentrations of this microorganism.

Factors that Diminish the Relationship Between Indicators and Pathogens
Seasonality of certain pathogens
Rate of infection in the host reservoir (herd, human population)
Differential decay
Differences in transport
Differences in sedimentation rates and partitioning to soil, sand, and sediments

Epidemiological Studies

Epidemiology studies have been conducted around the world to understand the correlative relationship between indicator concentrations and human health. The studies that have been conducted to date, and their methodologies, are summarized by Boehm and Soller (see ► Recreational Water Risk: Pathogens and Fecal Indicators. Most of the studies have focused on the health effects of recreational exposure to human fecal contamination from publicly owned treatment work discharges. These studies generally show a statistically significant correlation between enterococci and GI illness [256] in marine waters and *E. coli* and GI illness at freshwater beaches. Epidemiology studies are the cornerstone of the USA and EU water quality criteria and directives [31]. Acceptable illness rates are anchored to concentrations of indicator organisms in order to set acceptable contaminant levels. In the USA, 19 illnesses per 1,000 people is the acceptable illness level for marine water recreation, and in freshwater, the acceptable level is 8 illnesses per 1,000 people.

There are several important knowledge gaps in the understanding of how fecal contamination in recreational waters affects human health [31]. Few studies have documented the human health effects from exposure to nonhuman sources of fecal contamination including, but not limited to, bird and dog feces and urban and agricultural runoff [52, 98, 156]. A review of these studies suggests the relationship between indicator concentration and recreational waterborne illness risks is equivocal. Current studies with QMRA are trying to more fully understand the risks for exposure to animal feces in recreational waters [233].

Fecal Indicator Applications. There are numerous applications for fecal indicators and indicators need to be chosen that best serve a specific purpose or goal. One primary purpose of an indicator is to evaluate the public health risk for recreational water. In this case, general indicators may be employed since beach managers will need to know if fecal pollution is present and at what level. Since the presence of pathogens is highly dependent on the source of fecal pollution, adequate protection of public health will depend on assuming that the indicators are derived from sources that carry the highest pathogen burden. Rapid detection of a fecal indicator is more important than the level of information provided by the indicator since water quality can change rapidly in the beach environment [29]. Ultimately, the source of fecal pollution needs to be identified and remediated to remove the health risk.

Fecal indicators also serve as important tools for sanitary survey practices and for prioritizing remediation strategies. While daily monitoring with a general indicator such as enterococci or *E. coli* will provide information on the extent of fecal pollution, the source needs to be identified in order to take corrective actions. Both extensive mapping of the physical location of fecal pollution inputs (where is it coming from?) and determination of the host sources (is it human or nonhuman sources?) are necessary. Host-associated alternative indicators are best suited for these applications.

Source identifiers can also be used to evaluate the success of best management practices and influence of many green infrastructure efforts in agriculture and urban run-off settings. For example, the installation of tile drainage systems or constructed wetlands is commonly used to control the flow of agricultural waste across the landscape during rain events. Host-associated methods provide an excellent metric for estimating the efficiency of these waste management practices.

Rapid Methods for Indicators

Recreational water quality monitoring has traditionally relied upon culture-based methods and therefore test results are not available to the public until, at the earliest, the following day. It is well established that water quality can change in a matter of hours [29, 266]. A high priority for beach managers is to utilize rapid testing methods, many of which are based on qPCR of fecal indicators. Studies have compared different rapid methods [93]. New water quality criteria that are being formulated by the USEPA are expected to include rapid methods for enterococci using qPCR.

BEACH Act Legislation. The Beaches Environmental Assessment and Coastal Health (BEACH) Act of 2000 is an amendment to the Federal Water Pollution Control Act (commonly known as the Clean Water Act). This legislation required states and tribes to adopt new or revised water quality standards by 2004. It also required the USEPA to publish new or revised criteria for pathogens and pathogen indicators. The BEACH Act authorized appropriations for states and tribes to develop and implement water quality monitoring and public notification programs at recreational beaches. The USEPA has identified scientific gaps that need to be filled in order to develop improved water quality criteria [251].

Future Directions

The identification of host-associated source identifiers represents the first step toward the successful implementation of a fecal indicator method. Several additional steps must be taken to complete the method development phase including method optimization, design of appropriate laboratory controls, and defining a data interpretation model. After method development, it is necessary to define the operational parameters of the method. In the case of qPCR, this might include factors such as generation of a calibration curve, defining the range of quantification, precision, and limit of detection. The next step is to characterize the robustness of the method by measuring specificity, host distribution of the source identifier, abundance of source identifier in target group, describing fate and transport mechanisms, establishing links to general fecal indicators, pathogens, and public health outcomes. Once the operational parameters and robustness of the method are adequately described, a multiple laboratory validation study should be conducted to address issues of reproducibility, variability between laboratories, normalization of results, standardization of controls, minimum requirements to establish laboratory efficiency, and requirements for laboratory training. It is important to note that this list is not comprehensive. There may be additional steps required depending on the intended use of the method.

Rapidly advancing technologies will provide new opportunities to expand the number and types of fecal indicators. Next-generation sequencing technologies have increased our capacity to analyze whole microbial communities, rather than single organisms. Advancing technologies will also allow for more detailed analyses of the dynamics of fecal indicators in the environment. Further, more sensitive, specific, and rapid detection strategies are needed to improve monitoring programs for devising pollution remediation strategies and for the protection of public health.

Acknowledgments

We would like to thank Dr. Alford Dufour for insightful discussion and for providing historical information on the use of traditional indicators and evolution of water quality criteria. The US Environmental Protection Agency, through its Office of Research and Development, funded and managed, or partially funded and collaborated in, the research described herein. It has been subjected to the Agency's peer and administrative review and has been approved for external publication. Any opinions expressed in this entry are those of the author(s) and do not necessarily reflect the views of the Agency; therefore, no official endorsement should be inferred. Any mention of trade names or commercial products does not constitute endorsement or recommendation for use.

Bibliography

1. Ahmed W, Powell D, Goonetilleke A, Gardner T (2008) Detection and source identification of faecal pollution in non-sewered catchment by means of host-specific molecular markers. Water Sci Technol 58:579–586
2. Ahmed W, Stewart J, Gardner T, Powell D (2008) A real-time polymerase chain reaction assay for quantitative detection of the human-specific enterococci surface protein marker in sewage and environmental waters. Environ Microbiol 10:3255–3264
3. Ahmed W, Stewart J, Powell D, Gardner T (2008) Evaluation of Bacteroides markers for the detection of human faecal pollution. Lett Appl Microbiol 46:237–242
4. Ahmed W, Goonetilleke A, Powell D, Chauhan K, Gardner T (2009) Comparison of molecular markers to detect fresh sewage in environmental waters. Water Res 43:4908–4917
5. Ahmed W, Sawant S, Huygens F, Goonetilleke A, Gardner T (2009) Prevalence and occurrence of zoonotic bacterial pathogens in surface waters determined by quantitative PCR. Water Res 43:4918–4928
6. Ahmed W, Wan C, Goonetilleke A, Gardner T (2010) Evaluating sewage-associated JCV and BKV polyomaviruses for sourcing human fecal pollution in a coastal river in Southeast Queensland, Australia. J Environ Qual 39:1743–1750
7. Alm EW, Burke J, Spain A (2003) Fecal indicator bacteria are abundant in wet sand at freshwater beaches. Water Res 37:3978–3982
8. Alonso JL, Alonso MA (1993) Presence of *Campylobacter* in marine waters of Valencia. Spain Water Res 27:1559–1562
9. Anderson KL, Whitlock JE, Harwood VJ (2005) Persistence and differential survival of fecal indicator bacteria in subtropical waters and sediments. Appl Environ Microbiol 71:3041–3048
10. Anderson MA, Whitlock JE, Harwood VJ (2006) Diversity and distribution of *Escherichia coli* genotypes and antibiotic resistance phenotypes in feces of humans, cattle, and horses. Appl Environ Microbiol 72:6914–6922
11. Arai T, Ikejima N, Itoh T, Sakai S, Shimada T, Sakazaki R (1980) A survey of Plesiomonas shigelloides from aquatic environments, domestic animals, pets and humans. J Hyg (Lond) 84:203–211
12. Araujo RM, Puig A, Lasobras J, Lucena F, Jofre J (1997) Phages of enteric bacteria in fresh water with different levels of faecal pollution. J Appl Microbiol 82:281–286
13. Arnone RD, Walling JP (2007) Waterborne pathogens in urban watersheds. J Water Health 5:149–162
14. Ashbolt NJ, Grabow OK, Snozzi M (2001) Indicators of microbial water quality. In: Fewtrell L, Bartram J (eds) Water quality: guidelines, standards, and health. IWA Publishing, London
15. Ashbolt NJ, Schoen ME, Soller JA, Roser DJ (2010) Predicting pathogen risks to aid beach management: the real value of quantitative microbial risk assessment (QMRA). Water Res 44:4692–4703
16. Association, A. P. H. (1999) Standard methods for the examination of water and wastewater, 18th edn. American Public Health Association, Washington, DC
17. Aw TG, Gin KY, Ean Oon LL, Chen EX, Woo CH (2009) Prevalence and genotypes of human noroviruses in tropical urban surface waters and clinical samples in Singapore. Appl Environ Microbiol 75:4984–4992
18. Bae S, Wuertz S (2009) Rapid decay of host-specific fecal Bacteroidales cells in seawater as measured by quantitative PCR with propidium monoazide. Water Res 43:4850–4859
19. Bai S, Lung WS (2005) Modeling sediment impact on the transport of fecal bacteria. Water Res 39:5232–5240
20. Balleste E, Bonjoch X, Belanche LA, Blanch AR (2010) Molecular indicators used in the development of predictive models for microbial source tracking. Appl Environ Microbiol 76:1789–1795
21. Baudart J, Lemarchand K, Brisabois A, Lebaron P (2000) Diversity of *Salmonella* strains isolated from the aquatic environment as determined by serotyping and amplification of the ribosomal DNA spacer regions. Appl Environ Microbiol 66:1544–1552
22. Baums IB, Goodwin KD, Kiesling T, Wanless D, Diaz MR, Fell JW (2007) Luminex detection of fecal indicators in river samples,

marine recreational water, and beach sand. Mar Pollut Bull 54:521–536

23. Bell A, Layton AC, McKay L, Williams D, Gentry R, Sayler GS (2009) Factors influencing the persistence of fecal Bacteroides in stream water. J Environ Qual 38:1224–1232
24. Bernhard AE, Field KG (2000) Identification of nonpoint sources of fecal pollution in coastal waters by using host-specific 16 S ribosomal DNA genetic markers from fecal anaerobes. Appl Environ Microbiol 66:1587–1594
25. Bernhard AE, Field KG (2000) A PCR assay to discriminate human and ruminant feces on the basis of host differences in Bacteroides-Prevotella genes encoding 16 S rRNA. Appl Environ Microbiol 66:4571–4574
26. Beversdorf LJ, Bornstein-Forst SM, McLellan SL (2007) The potential for beach sand to serve as a reservoir for *Escherichia coli* and the physical influences on cell die-off. J Appl Microbiol 102:1372–1381
27. Bienfang PK, Defelice SV, Laws EA, Brand LE, Bidigare RR, Christensen S, Trapido-Rosenthal H, Hemscheidt TK, McGillicuddy DJ, Anderson DM, Solo-Gabriele HM, Boehm AB, Backer LC (2011) Prominent human health impacts from several marine microbes: history, ecology, and public health implications. Int J Microbiol 2011:152815
28. Blanch AR, Belanche-Munoz L, Bonjoch X, Ebdon J, Gantzer C, Lucena F, Ottoson J, Kourtis C, Iversen A, Kuhn I, Moce L, Muniesa M, Schwartzbrod J, Skraber S, Papageorgiou GT, Taylor H, Wallis J, Jofre J (2006) Integrated analysis of established and novel microbial and chemical methods for microbial source tracking. Appl Environ Microbiol 72:5915–5926
29. Boehm AB (2007) Enterococci concentrations in diverse coastal environments exhibit extreme variability. Environ Sci Technol 41:8227–8232
30. Boehm AB, Keymer DP, Shellenbarger GG (2005) An analytical model of enterococci inactivation, grazing, and transport in the surf zone of a marine beach. Water Res 39:3565–3578
31. Boehm AB, Ashbolt NJ, Colford JM Jr, Dunbar LE, Fleming LE, Gold MA, Hansel JA, Hunter PR, Ichida AM, McGee CD, Soller JA, Weisberg SB (2009) A sea change ahead for recreational water quality criteria. J Water Health 7:9–20
32. Boehm AB, Yamahara KM, Love DC, Peterson BM, McNeill K, Nelson KL (2009) Covariation and photoinactivation of traditional and novel indicator organisms and human viruses at a sewage-impacted marine beach. Environ Sci Technol 43:8046–8052
33. Bolton F, Surman SB, Martin K, Wareing DR, Humphrey TJ (1999) Presence of *Campylobacter* and *Salmonella* in sand from bathing beaches. Epidemiol Infect 122:7–13
34. Bonde GJ (1963) Bacterial indicators of water pollution. A study of quantitative estimation. Teknisk Forlag, Copenhagen
35. Bonjoch X, Balleste E, Blanch AR (2004) Multiplex PCR with 16 S rRNA gene-targeted primers of bifidobacterium spp. to identify sources of fecal pollution. Appl Environ Microbiol 70:3171–3175
36. Bothner MH, Takada H, Knight IT, Hill RT, Butman B, Farrington JW, Colwell RR, Grassle JF (1994) Sewage contamination in sediments beneath a deep-ocean dump site off New-York. Mar Environ Res 38:43–59
37. Bower PA, Scopel CO, Jensen ET, Depas MM, McLellan SL (2005) Detection of genetic markers of fecal indicator bacteria in Lake Michigan and determination of their relationship to *Escherichia coli* densities using standard microbiological methods. Appl Environ Microbiol 71:8305–8313
38. Brion GM, Meschke JS, Sobsey MD (2002) F-specific RNA coliphages: occurrence, types, and survival in natural waters. Water Res 36:2419–2425
39. Cabelli VJ, Dufour AP, McCabe LJ, Levin MA (1982) Swimming-associated gastroenteritis and water quality. Am J Epidemiol 115:606–616
40. Caldwell JM, Raley ME, Levine JF (2007) Mitochondrial multiplex real-time PCR as a source tracking method in fecal-contaminated effluents. Environ Sci Technol 41:3277–3283
41. Cao Y, Griffith JF, Weisberg SB (2009) Evaluation of optical brightener photodecay characteristics for detection of human fecal contamination. Water Res 43:2273–2279
42. Caplenas NR, Kanarek MS (1984) Thermotolerant non-fecal source Klebsiella pneumoniae: validity of the fecal coliform test in recreational waters. Am J Public Health 74:1273–1275
43. Carson CA, Shear BL, Ellersieck MR, Asfaw A (2001) Identification of fecal *Escherichia coli* from humans and animals by ribotyping. Appl Environ Microbiol 67:1503–1507
44. Carson CA, Shear BL, Ellersieck MR, Schnell JD (2003) Comparison of ribotyping and repetitive extragenic palindromic-PCR for identification of fecal *Escherichia coli* from humans and animals. Appl Environ Microbiol 69:1836–1839
45. Carson CA, Christiansen JM, Yampara-Iquise H, Benson VW, Baffaut C, Davis JV, Broz RR, Kurtz WB, Rogers WM, Fales WH (2005) Specificity of a Bacteroides thetaiotaomicron marker for human feces. Appl Environ Microbiol 71:4945–4949
46. Center for Disease Control (2011) posting date. *Escherichia coli* O157:H7, General information
47. Chapron CD, Ballester NA, Fontaine JH, Frades CN, Margolin AB (2000) Detection of astroviruses, enteroviruses, and adenovirus types 40 and 41 in surface waters collected and evaluated by the information collection rule and an integrated cell culture-nested PCR procedure. Appl Environ Microbiol 66:2520–2525
48. Charoenca N, Fujioka RS (1993) Assessment of *Staphylococcus* bacteria in Hawaii's marine recreational waters. Water Sci Technol 27:283–289
49. Charoenca N, Fujioka RS (1995) Association of staphylococcal skin infections and swimming. Water Sci Technol 31:11–17
50. Chern EC, Tsai YL, Olson BH (2004) Occurrence of genes associated with enterotoxigenic and enterohemorrhagic *Escherichia coli* in agricultural waste lagoons. Appl Environ Microbiol 70:356–362
51. Choi S, Jiang SC (2005) Real-time PCR quantification of human adenoviruses in urban rivers indicates genome prevalence but low infectivity. Appl Environ Microbiol 71:7426–7433

52. Colford JM Jr, Wade TJ, Schiff KC, Wright CC, Griffith JF, Sandhu SK, Burns S, Sobsey M, Lovelace G, Weisberg SB (2007) Water quality indicators and the risk of illness at beaches with nonpoint sources of fecal contamination. Epidemiology 18:27–35
53. Converse RR, Blackwood AD, Kirs M, Griffith JF, Noble RT (2009) Rapid QPCR-based assay for fecal Bacteroides spp. as a tool for assessing fecal contamination in recreational waters. Water Res 43:4828–4837
54. Davies CM, Long JA, Donald M, Ashbolt NJ (1995) Survival of fecal microorganisms in marine and freshwater sediments. Appl Environ Microbiol 61:1888–1896
55. DePaola A, Hopkins LH, Peeler JT, Wentz B, McPhearson RM (1990) Incidence of *Vibrio parahaemolyticus* in U.S. coastal waters and oysters. Appl Environ Microbiol 56:2299–2302
56. Desmarais TR, Solo-Gabriele HM, Palmer CJ (2002) Influence of soil on fecal indicator organisms in a tidally influenced subtropical environment. Appl Environ Microbiol 68:1165–1172
57. Dethlefsen L, Huse S, Sogin ML, Relman DA (2008) The pervasive effects of an antibiotic on the human gut microbiota, as revealed by deep 16 S rRNA sequencing. PLoS Biol 6:e280
58. Devriese LLA, van de Kerckhove A, Kilpper-Baelz R, Schleifer K (1987) Characterization and identification of *Enterococcus* species isolated from the intestines of animals. Int J Syst Bacteriol 37:257–259
59. Dick LK, Field KG (2004) Rapid estimation of numbers of fecal Bacteroidetes by use of a quantitative PCR assay for 16 S rRNA genes. Appl Environ Microbiol 70:5695–5697
60. Dick LK, Bernhard AE, Brodeur TJ, Santo Domingo JW, Simpson JM, Walters SP, Field KG (2005) Host distributions of uncultivated fecal Bacteroidales bacteria reveal genetic markers for fecal source identification. Appl Environ Microbiol 71:3184–3191
61. Dick LK, Stelzer EA, Bertke EE, Fong DL, Stoeckel DM (2010) Relative decay of Bacteroidales microbial source tracking markers and cultivated *Escherichia coli* in freshwater microcosms. Appl Environ Microbiol 76:3255–3262
62. Dickerson JW Jr, Hagedorn C, Hassall A (2007) Detection and remediation of human-origin pollution at two public beaches in Virginia using multiple source tracking methods. Water Res 41:3758–3770
63. Dombek PE, Johnson LK, Zimmerley ST, Sadowsky MJ (2000) Use of repetitive DNA sequences and the PCR To differentiate *Escherichia coli* isolates from human and animal sources. Appl Environ Microbiol 66:2572–2577
64. Dorai-Raj S, O'Grady J, Colleran E (2009) Specificity and sensitivity evaluation of novel and existing Bacteroidales and Bifidobacteria-specific PCR assays on feces and sewage samples and their application for microbial source tracking in Ireland. Water Res 43:4980–4988
65. Dufour AP (1984) Bacterial indicators of recreational water quality. Can J Public Health 75:49–56
66. Dufour AP, Schaub S (2007) The evolution of water quality criteria in the United Sates, 1922–2003. In: Wymer LJ (ed) Statistical framework for recreational water quality monitoring. Wiley, New York
67. Eckburg PB, Bik EM, Bernstein CN, Purdom E, Dethlefsen L, Sargent M, Gill SR, Nelson KE, Relman DA (2005) Diversity of the human intestinal microbial flora. Science 308:1635–1638
68. Eckner KF (1998) Comparison of membrane filtration and multiple-tube fermentation by the colilert and enterolert methods for detection of waterborne coliform bacteria, *Escherichia coli*, and enterococci used in drinking and bathing water quality monitoring in southern Sweden. Appl Environ Microbiol 64:3079–3083
69. Edge TA, Boehm AB (2011) Classical and molecular methods to measure fecal indicator bacteria. In: Sadowsky MJ, Whitman RL (eds) The fecal indicator bacteria. ASM Press, Washington, DC
70. Edwards DD, McFeters GA, Venkatesan MI (1998) Distribution of *Clostridium perfringens* and fecal sterols in a benthic coastal marine environment influenced by the sewage outfall from McMurdo Station, Antarctica. Appl Environ Microbiol 64:2596–2600
71. Emerson DJ, Cabelli VJ (1982) Extraction of *Clostridium perfringens* spores from bottom sediment samples. Appl Environ Microbiol 44:1144–1149
72. Field KG, Samadpour M (2007) Fecal source tracking, the indicator paradigm, and managing water quality. Water Res 41:3517–3538
73. Field KG, Bernhard AE, Brodeur TJ (2003) Molecular approaches to microbiological monitoring: fecal source detection. Environ Monit Assess 81:313–326
74. Fiksdal L, Maki JS, LaCroix SJ, Staley JT (1985) Survival and detection of Bacteroides spp., prospective indicator bacteria. Appl Environ Microbiol 49:148–150
75. Fogarty LR, Voytek MA (2005) Comparison of bacteroides-prevotella 16 S rRNA genetic markers for fecal samples from different animal species. Appl Environ Microbiol 71:5999–6007
76. Fong T-T, Lipp EK (2005) Enteric viruses of humans and animals in aquatic environments: health risks, detection, and potential water quality assessment tools. Microbiol Mol Biol Rev 69:357–371
77. Frank DN, St Amand AL, Feldman RA, Boedeker EC, Harpaz N, Pace NR (2007) Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases. Proc Natl Acad Sci USA 104:13780–13785
78. Fremaux B, Gritzfeld J, Boa T, Yost CK (2009) Evaluation of host-specific Bacteroidales 16 S rRNA gene markers as a complementary tool for detecting fecal pollution in a prairie watershed. Water Res 43:4838–4849
79. Fujioka RS (2001) Monitoring coastal marine waters for spore-forming bacteria of faecal and soil origin to determine point from non-point source pollution. Water Sci Technol 44:181–188
80. Fung DYC, Fujioka R, Vijayavel K, Sato D, Bishop D (2007) Evaluation of Fung double tube test for *Clostridium perfringens* and Easyphage test for F-specific RNA coliphages as rapid screening tests for fecal contamination in recreational waters of Hawaii (vol 15, pg 217, 2007). J Rapid Meth Aut Mic 15:411–411
81. Furet JP, Firmesse O, Gourmelon M, Bridonneau C, Tap J, Mondot S, Dore J, Corthier G (2009) Comparative assessment

of human and farm animal faecal microbiota using real-time quantitative PCR. FEMS Microbiol Ecol 68:351–362

82. Gast RJ, Moran D, Dennett MR, Wurtsbaugh WA, Amaral-Zettler LA (2011) Amoebae and *Legionella pneumophila* in saline environments. J Water Health 9:37–52
83. Gauthier F, Neufeld JD, Driscoll BT, Archibald FS (2000) Coliform bacteria and nitrogen fixation in pulp and paper mill effluent treatment systems. Appl Environ Microbiol 66:5155–5160
84. Gawler AH, Beecher JE, Brandao J, Carroll NM, Falcao L, Gourmelon M, Masterson B, Nunes B, Porter J, Rince A, Rodrigues R, Thorp M, Walters JM, Meijer WG (2007) Validation of host-specific Bacteriodales 16 S rRNA genes as markers to determine the origin of faecal pollution in Atlantic Rim countries of the European Union. Water Res 41:3780–3784
85. Gerba CP, Rose JB, Haas CN, Crabtree KD (1996) Waterborne rotavirus: a risk assessment. Water Res 30:2929–2940
86. Gersberg RM, Rose MA, Robles-Sikisaka R, Dhar AK (2006) Quantitative detection of hepatitis A virus and enteroviruses near the United States-Mexico border and correlation with levels of fecal indicator bacteria. Appl Environ Microbiol 72:7438–7444
87. Gilpin B, James T, Nourozi F, Saunders D, Scholes P, Savill M (2003) The use of chemical and molecular microbial indicators for faecal source identification. Water Sci Technol 47:39–43
88. Goodwin KD, Pobuda M (2009) Performance of CHROMagar Staph aureus and CHROMagar MRSA for detection of *Staphylococcus aureus* in seawater and beach sand – comparison of culture agglutination, and molecular analyses. Water Res 43:4802–4811
89. Gourmelon M, Caprais MP, Mieszkin S, Marti R, Wery N, Jarde E, Derrien M, Jadas-Hecart A, Communal PY, Jaffrezic A, Pourcher AM (2010) Development of microbial and chemical MST tools to identify the origin of the faecal pollution in bathing and shellfish harvesting waters in France. Water Res 44:4812–4824
90. Graczyk TK, Sunderland D, Tamang L, Lucy FE, Breysse PN (2007) Bather density and levels of *Cryptosporidium* Giardia, and pathogenic microsporidian spores in recreational bathing water. Parasitol Res 101:1729–1731
91. Griffin DW, Gibson CJ 3rd, Lipp EK, Riley K, Paul JH 3rd, Rose JB (1999) Detection of viral pathogens by reverse transcriptase PCR and of microbial indicators by standard methods in the canals of the Florida Keys. Appl Environ Microbiol 65:4118–4125
92. Griffin DW, Lipp EK, McLaughlin MR, Rose JB (2001) Marine recreation and public health microbiology: quest for the ideal indicator. Bioscience 51:817–825
93. Griffith JF, Cao Y, McGee CD, Weisberg SB (2009) Evaluation of rapid methods and novel indicators for assessing microbiological beach water quality. Water Res 43:4900–4907
94. Grimes DJ (1975) Release of sediment-bound fecal coliforms by dredging. Appl Microbiol 29:109–111
95. Haack SK, Duris JW, Fogarty LR, Kolpin DW, Focazio MJ, Furlong ET, Meyer MT (2009) Comparing wastewater chemicals, indicator bacteria concentrations, and bacterial pathogen genes as fecal pollution indicators. J Environ Qual 38:248–258
96. Haake DA, Dundoo M, Cader R, Kubak BM, Hartskeerl RA, Sejvar JJ, Ashford DA (2002) Leptospirosis, water sports, and chemoprophylaxis. Clin Infect Dis 34:E40–E43
97. Hagedorn C, Robinson SL, Filtz JR, Grubbs SM, Angier TA, Reneau RB Jr (1999) Determining sources of fecal pollution in a rural Virginia watershed with antibiotic resistance patterns in fecal streptococci. Appl Environ Microbiol 65: 5522–5531
98. Haile RW, Witte JS, Gold M, Cressey R, McGee C, Millikan RC, Glasser A, Harawa N, Ervin C, Harmon P, Harper J, Dermand J, Alamillo J, Barrett K, Nides M, Wang G (1999) The health effects of swimming in ocean water contaminated by storm drain runoff. Epidemiology 10:355–363
99. Halbur PG, Kasorndorkbua C, Gilbert C, Guenette D, Potters MB, Purcell RH, Emerson SU, Toth TE, Meng XJ (2001) Comparative pathogenesis of infection of pigs with hepatitis E viruses recovered from a pig and a human. J Clin Microbiol 39:918–923
100. Haley BJ, Cole DJ, Lipp EK (2009) Distribution, diversity, and seasonality of waterborne salmonellae in a rural watershed. Appl Environ Microbiol 75:1248–1255
101. Hamilton MJ, Yan T, Sadowsky MJ (2006) Development of goose- and duck-specific DNA markers to determine sources of *Escherichia coli* in waterways. Appl Environ Microbiol 72:4012–4019
102. Harrison S, Kinra S (2004) Outbreak of *Escherichia coli* O157 associated with a busy beach. Commun Dis Public Health 7:47–50
103. Hartel PG, Summer JD, Hill JL, Collins JV, Entry JA, Segars WI (2002) Geographic variability of *Escherichia coli* ribotypes from animals in Idaho and Georgia. J Environ Qual 31: 1273–1278
104. Harvey S, Greenwood JR, Pickett MJ, Mah RA (1976) Recovery of *Yersinia enterocolitica* from streams and lakes of California. Appl Environ Microbiol 32:352–354
105. Harwood VJ, Whitlock J, Withington V (2000) Classification of antibiotic resistance patterns of indicator bacteria by discriminant analysis: use in predicting the source of fecal contamination in subtropical waters. Appl Environ Microbiol 66: 3698–3704
106. Haugland RA, Siefring SC, Wymer LJ, Brenner KP, Dufour AP (2005) Comparison of Enterococcus measurements in freshwater at two recreational beaches by quantitative polymerase chain reaction and membrane filter culture analysis. Water Res 39:559–568
107. Hayashi H, Sakamoto M, Kitahara M, Benno Y (2006) Diversity of the Clostridium coccoides group in human fecal microbiota as determined by 16 S rRNA gene library. FEMS Microbiol Lett 257:202–207
108. He J, Jiang S (2005) Quantification of enterococci and human adenoviruses in environmental samples by real-time PCR. Appl Environ Microbiol 71:2250–2255
109. Higgins JA, Belt KT, Karns JS, Russell-Anelli J, Shelton DR (2005) *tir*- and *stx*-positive *Escherichia coli* in stream waters in a metropolitan area. Appl Environ Microbiol 71:2511–2519

110. Hill RT, Straube WL, Palmisano AC, Gibson SL, Colwell RR (1996) Distribution of sewage indicated by *Clostridium perfringens* at a deep-water disposal site after cessation of sewage disposal. Appl Environ Microbiol 62:1741–1746

111. Hipsey MR, Antenucci JP, Brookes JD (2008) A generic, process-based model of microbial pollution in aquatic systems. Water Resour Res 44:26

112. Horman A, Rimhanen-Finne R, Maunula L, von Bonsdorff C-H, Torvela N, Heikinheimo A, Hanninen M-L (2004) *Campylobacter* spp., *Giardia* spp., *Cryptosporidium* spp., Noroviruses, and indicator organisms in surface water in southwestern Finland, 2000–2001. Appl Environ Microbiol 70:87–95

113. Hou D, Rabinovici SJM, Boehm AB (2006) Enterococci predictions from partial least squares regression models in conjunction with a single-sample standard improve the efficacy of beach management advisories. Environ Sci Technol 40:1737–1743

114. Hsu FC, Shieh YS, van Duin J, Beekwilder MJ, Sobsey MD (1995) Genotyping male-specific RNA coliphages by hybridization with oligonucleotide probes. Appl Environ Microbiol 61:3960–3966

115. Hussain MA, Ford R, Hill J (2010) Determination of fecal contamination indicator sterols in an Australian water supply system. Environ Monit Assess 165:147–157

116. International Organization for Standardization (2000) Water quality – Detection and enumeration of intestinal enterococci ISO 7899–2:000

117. Ishii S, Ksoll WB, Hicks RE, Sadowsky MJ (2006) Presence and growth of naturalized *Escherichia coli* in temperate soils from Lake Superior watersheds. Appl Environ Microbiol 72:612–621

118. Ishii S, Yan T, Shivley DA, Byappanahalli MN, Whitman RL, Sadowsky MJ (2006) *Cladophora* (Chlorophyta) spp. harbor human bacterial pathogens in nearshore water of Lake Michigan. Appl Environ Microbiol 72:4545–4553

119. Jamieson R, Joy DM, Lee H, Kostaschuk R, Gordon R (2005) Transport and deposition of sediment-associated *Escherichia coli* in natural streams. Water Res 39:2665–2675

120. Jamieson RC, Joy DM, Lee H, Kostaschuk R, Gordon RJ (2005) Resuspension of sediment-associated *Escherichia coli* in a natural stream. J Environ Qual 34:581–589

121. Jeter SN, McDermott CM, Bower PA, Kinzelman JL, Bootsma MJ, Goetz GW, McLellan SL (2009) Bacteroidales diversity in ring-billed gulls (Laurus delawarensis) residing at Lake Michigan beaches. Appl Environ Microbiol 75:1525–1533

122. Jiang SC, Chu W (2004) PCR detection of pathogenic viruses in southern California urban rivers. J Appl Microbiol 97:17–28

123. Jiang S, Noble R, Chu W (2001) Human adenoviruses and coliphages in urban runoff-impacted coastal waters of Southern California. Appl Environ Microbiol 67: 179–184

124. Jiang SC, Nobel R, Chu W (2001) Human adenoviruses and coliphage in urban runoff-impacted coastal waters of southern California. Appl Environ Microbiol 67:179–184

125. Jiang SC, Chu W, He JW (2007) Seasonal detection of human viruses and coliphage in Newport Bay, California. Appl Environ Microbiol 73:6468–6474

126. Jofre J, Blasi M, Bosch A, Lucena F (1989) Occurrence of bacteriophages infecting Bacteroides-Fragilis and other viruses in polluted marine-sediments. Water Sci Technol 21:15–19

127. Johnson CN, Flowers AR, Noriea NF III, Zimmerman AM, Bowers JC, DePaola A, Grimes DJ (2010) Relationships between environmental factors and pathogenic Vibrios in the Northern Gulf of Mexico. Appl Environ Microbiol 76:7076–7084

128. Johnston C, Ufnar JA, Griffith JF, Gooch JA, Stewart JR (2010) A real-time qPCR assay for the detection of the nifH gene of Methanobrevibacter smithii, a potential indicator of sewage pollution. J Appl Microbiol 109:1946–1956

129. Keymer DP, Miller MC, Schoolnik GK, Boehm AB (2007) Genomic and phenotypic diversity of coastal *Vibrio cholerae* is explained by environmental factors. Appl Environ Microbiol 73:3705–3714

130. Keymer DP, Lam L, Boehm AB (2009) Biogeographic patterns in genomic diversity among a large collection of *Vibrio cholerae* isolates. Appl Environ Microbiol 75:1658–1666

131. Khatib LA, Tsai YL, Olson BH (2002) A biomarker for the identification of cattle fecal pollution in water using the LTIIa toxin gene from enterotoxigenic *Escherichia coli*. Appl Microbiol Biotechnol 59:97–104

132. Khatib LA, Tsai YL, Olson BH (2003) A biomarker for the identification of swine fecal pollution in water, using the STII toxin gene from enterotoxigenic *Escherichia coli*. Appl Microbiol Biotechnol 63:231–238

133. Kildare BJ, Leutenegger CM, McSwain BS, Bambic DG, Rajal VB, Wuertz S (2007) 16 S rRNA-based assays for quantitative detection of universal, human-, cow-, and dog-specific fecal Bacteroidales: a Bayesian approach. Water Res 41:3701–3715

134. Kinzelman J, McLellan SL, Daniels AD, Cashin S, Singh A, Gradus S, Bagley R (2004) Non-point source pollution: determination of replication versus persistence of *Escherichia coli* in surface water and sediments with correlation of levels to readily measurable environmental parameters. J Water Health 2:103–114

135. Kirs M, Smith DC (2007) Multiplex quantitative real-time reverse transcriptase PCR for F+ −specific RNA coliphages: a method for use in microbial source tracking. Appl Environ Microbiol 73:808–814

136. Kreader CA (1995) Design and evaluation of Bacteroides DNA probes for the specific detection of human fecal pollution. Appl Environ Microbiol 61:1171–1179

137. Lamendella R, Domingo JW, Oerther DB, Vogel JR, Stoeckel DM (2007) Assessment of fecal pollution sources in a small northern-plains watershed using PCR and phylogenetic analyses of Bacteroidetes 16 S rRNA gene. FEMS Microbiol Ecol 59:651–660

138. Layton A, McKay L, Williams D, Garrett V, Gentry R, Sayler G (2006) Development of Bacteroides 16 S rRNA gene

TaqMan-based real-time PCR assays for estimation of total, human, and bovine fecal pollution in water. Appl Environ Microbiol 72:4214–4224

139. Layton BA, Walters SP, Lam LH, Boehm AB (2010) Enterococcus species distribution among human and animal hosts using multiplex PCR. J Appl Microbiol 109:539–547

140. Leach MD, Broschat SL, Call DR (2008) A discrete, stochastic model and correction method for bacterial source tracking. Environ Sci Technol 42:524–529

141. Leclerc H, Schwartzbrod L, Dei-Cas E (2002) Microbial agents associated with waterborne diseases. Crit Rev Microbiol 28:371–409

142. Lee CS, Lee J (2010) Evaluation of new gyrB-based real-time PCR system for the detection of *B. fragilis* as an indicator of human-specific fecal contamination. J Microbiol Meth 82:311–318

143. Lee DY, Weir SC, Lee H, Trevors JT (2010) Quantitative identification of fecal water pollution sources by TaqMan real-time PCR assays using *Bacteroidales* 16 S rRNA genetic markers. Appl Microbiol Biotechnol 88:1373–1383

144. Lemarchand K, Lebaron P (2003) Occurrence of *Salmonella* spp. and *Cryptosporidium* spp. in a French coastal watershed: relationship with fecal indicators. FEMS Microbiol Lett 218:203–209

145. Lemarchand K, Masson L, Brousseau R (2004) Molecular biology and DNA microarray technology for microbial quality monitoring of water. Crit Rev Microbiol 30:145–172

146. Leung HD, Chen G, Sharma K (2005) Effect of detached/re-suspended solids from sewer sediment on the sewage phase bacterial activity. Water Sci Technol 52:147–152

147. Lipp EK, Huq A, Colwell RR (2002) Effects of global climate on infectious disease: the cholera model. Clin Microbiol Rev 15:757–770

148. Lynch PA, Gilpin BJ, Sinton LW, Savill MG (2002) The detection of Bifidobacterium adolescentis by colony hybridization as an indicator of human faecal pollution. J Appl Microbiol 92:526–533

149. Mac Kenzie WR, Hoxie NJ, Proctor ME, Gradus MS, Blair KA, Peterson DE, Kazmierczak JJ, Addiss DG, Fox KR, Rose JB et al (1994) A massive outbreak in Milwaukee of cryptosporidium infection transmitted through the public water supply. N Engl J Med 331:161–167

150. Malinen E, Rinttila T, Kajander K, Matto J, Kassinen A, Krogius L, Saarela M, Korpela R, Palva A (2005) Analysis of the fecal microbiota of irritable bowel syndrome patients and healthy controls with real-time PCR. Am J Gastroenterol 100:373–382

151. Mallin MA, Williams KE, Esham EC, Lowe RP (2000) Effect of human development on bacteriological water quality in coastal watersheds. Ecol Appl 10:1047–1056

152. Marshall MM, Naumovitz D, Ortega Y, Sterling CR (1997) Waterborne protozoan pathogens. Clin Microbiol Rev 10:67–85

153. Matsuki T, Watanabe K, Fujimoto J, Miyamoto Y, Takada T, Matsumoto K, Oyaizu H, Tanaka R (2002) Development of 16 S rRNA-gene-targeted group-specific primers for the detection and identification of predominant bacteria in human feces. Appl Environ Microbiol 68:5445–5451

154. Matsuki T, Watanabe K, Fujimoto J, Kado Y, Takada T, Matsumoto K, Tanaka R (2004) Quantitative PCR with 16 S rRNA-gene-targeted species-specific primers for analysis of human intestinal bifidobacteria. Appl Environ Microbiol 70:167–173

155. Matsuki T, Watanabe K, Fujimoto J, Takada T, Tanaka R (2004) Use of 16 S rRNA gene-targeted group-specific primers for real-time PCR analysis of predominant bacteria in human feces. Appl Environ Microbiol 70:7220–7228

156. McBride GB, Salmond CE, Bandaranayake DR, Turner SJ, Lewis GD, Till DG (1998) Health effects of marine bathing in New Zealand. Int J Environ Health Res 8:173–189

157. McLaughlin MR, Rose JB (2006) Application of *Bacteroides fragilis* phage as an alternative indicator of sewage pollution in Tampa Bay, Florida. Estuar Coast 29:246–256

158. McLellan SL, Daniels AD, Salmore AK (2001) Clonal populations of thermotolerant Enterobacteriaceae in recreational water and their potential interference with fecal *Escherichia coli* counts. Appl Environ Microbiol 67:4934–4938

159. McLellan SL, Daniels AD, Salmore AK (2003) Genetic characterization of *Escherichia coli* populations from host sources of fecal pollution by using DNA fingerprinting. Appl Environ Microbiol 69:2587–2594

160. McLellan SL, Hollis EJ, Depas MM, Van Dyke M, Harris J, Scopel CO (2007) Distribution and fate of *Escherichia coli* in Lake Michigan following contamination with urban stormwater and combined sewer overflows. J Great Lakes Res 33:566–580

161. McLellan SL, Huse SM, Mueller-Spitz SR, Andreishcheva EN, Sogin ML (2010) Diversity and population structure of sewage-derived microorganisms in wastewater treatment plant influent. Environ Microbiol 12:378–392

162. McQuaig SM, Scott TM, Harwood VJ, Farrah SR, Lukasik JO (2006) Detection of human-derived fecal pollution in environmental waters by use of a PCR-based human polyomavirus assay. Appl Environ Microbiol 72:7567–7574

163. McQuaig SM, Scott TM, Lukasik JO, Paul JH, Harwood VJ (2009) Quantification of human polyomaviruses JC Virus and BK Virus by TaqMan quantitative PCR and comparison to other water quality indicators in water and fecal samples. Appl Environ Microbiol 75:3379–3388

164. Medema GJ, Bahar M, Schets FM (1997) Survival of *Cryptosporidium parvum, Escherichia coli*, faecal enterococci and *Clostridium perfringens* in river water: influence of temperature and autochthonous microorganisms. Water Sci Technol 35:249–252

165. Mieszkin S, Yala JF, Joubrel R, Gourmelon M (2010) Phylogenetic analysis of *Bacteroidales* 16 S rRNA gene sequences from human and animal effluents and assessment of ruminant faecal pollution by real-time PCR. J Appl Microbiol 108:974–984

166. Miller WA, Miller MA, Gardner IA, Atwill ER, Byrne BA, Jang S, Harris M, Ames J, Jessup D, Paradies D, Worcester K, Melli A,

Conrad PA (2006) *Salmonella* spp., *Vibrio* spp., *Clostridium perfringens*, and *Plesiomonas shigelloides* in marine and freshwater invertebrates from coastal California ecosystems. Microb Ecol 52:198–206

167. Mocé-Llivina L, Lucena F, Jofre J (2005) Enteroviruses and bacteriophages in bathing waters. Appl Environ Microbiol 71:6838–6844
168. Moe C (2002) Waterborne transmission of infectious agents. In: Hurst CJ, Crawfod RL, Knudsen GR, McInerney MJ, Stetzenbach LD (eds) Manual of environmental microbiology, 2nd edn. ASM Press, Washington, DC, pp 184–204
169. Mueller-Spitz SR, Stewart LB, Klump JV, McLellan SL (2010) Freshwater suspended sediments and sewage are reservoirs for enterotoxin-positive *Clostridium perfringens*. Appl Environ Microbiol 76:5556–5562
170. Mumy KL, Findlay RH (2004) Convenient determination of DNA extraction efficiency using an external DNA recovery standard and quantitative-competitive PCR. J Microbiol Methods 57:259–268
171. Muniesa M, Jofre J, Garcia-Aljaro C, Blanch AR (2006) Occurrence of *Escherichia coli* O157:H7 and other enterohemorrhagic *Escherichia coli* in the environment. Environ Sci Technol 40:7141–7149
172. Muniesa M, Payan A, Moce-Llivina L, Blanch AR, Jofre J (2009) Differential persistence of F-specific RNA phage subgroups hinders their use as single tracers for faecal source tracking in surface water. Water Res 43:1559–1564
173. Nebra Y, Bonjoch X, Blanch AR (2003) Use of *Bifidobacterium dentium* as an indicator of the origin of fecal water pollution. Appl Environ Microbiol 69:2651–2656
174. Nevers MB, Boehm AB (2010) Modeling fate and transport of fecal bacteria in surface water. In: Sadowsky MJ, Whitman RL (eds) The fecal indicator bacteria. ASM Press, Washington, DC
175. Newton RJ, VandeWalle JL, Borchardt MA, Gorelick MH, McLellan SL (2011) *Lachnospiraceae* and *Bacteroidales* alternative fecal indicators reveal chronic human sewage contamination in an urban harbor. Appl Environ Microbiol 77:6972–6981
176. Nichols PD, Leeming R, Rayner MS, Latham V, Ashbolt NJ, Turner C (1993) Comparison of the abundance of the fecal sterol soprostanol and fecal bacerial groups in inner-shlef waters and sediments near Sydney, Australia. J Chromatogr 643:189–195
177. Niemela SI, Vaatanen P (1982) Survival in lake water of *Klebsiella pneumoniae* discharged by a paper mill. Appl Environ Microbiol 44:264–269
178. Noble RT, Fuhrman JD (2001) Enteroviruses detected by reverse transcriptase polymerase chain reaction from the coastal waters of Santa Monica Bay, California: low correlation to bacterial indicator levels. Hydrobiologia 460:175–183
179. Noble RT, Allen SM, Blackwood AD, Chu W, Jiang SC, Lovelace GL, Sobsey MD, Stewart JR, Wait DA (2003) Use of viral pathogens and indicators to differentiate between human and non-human fecal contamination in a microbial source tracking comparison study. J Water Health 1:195–207
180. Noble RT, Moore DF, Leecaster MK, McGee CD, Weisberg SB (2003) Comparison of total coliform, fecal coliform, and enterococcus bacterial indicator response for ocean recreational water quality testing. Water Res 37:1637–1643
181. Noble RT, Griffith JF, Blackwood AD, Fuhrman JA, Gregory JB, Hernandez X, Liang X, Bera AA, Schiff K (2006) Multitiered approach using quantitative PCR to track sources of fecal pollution affecting Santa Monica Bay, California. Appl Environ Microbiol 72:1604–1612
182. Noble RT, Blackwood AD, Griffith JF, McGee CD, Weisberg SB (2010) Comparison of rapid quantitative PCR-based and conventional culture-based methods for enumeration of *Enterococcus* spp. and *Escherichia coli* in recreational waters. Appl Environ Microbiol 76:7437–7443
183. Obiri-Danso K, Jones K (2000) Intertidal sediments as reservoirs for hippurate negative campylobacters, salmonellae and faecal indicators in three EU recognized bathing waters in North West England. Water Res 34:519–527
184. Okabe S, Shimazu Y (2007) Persistence of host-specific *Bacteroides-Prevotella* 16 S rRNA genetic markers in environmental waters: effects of temperature and salinity. Appl Microbiol Biotechnol 76:935–944
185. Okabe S, Okayama N, Savichtcheva O, Ito T (2007) Quantification of host-specific Bacteroides-Prevotella 16 S rRNA genetic markers for assessment of fecal pollution in freshwater. Appl Microbiol Biotechnol 74:890–901
186. Palmer CJ, Tsai YL, Paszko-Kolva C, Mayer C, Sangermano LR (1993) Detection of *Legionella* species in sewage and ocean water by polymerase chain reaction, direct fluorescent-antibody, and plate culture methods. Appl Environ Microbiol 59:3618–3624
187. Panicker G, Myers ML, Bej AK (2004) Rapid detection of *Vibrio vulnificus* in shellfish and Gulf of Mexico water by real-time PCR. Appl Environ Microbiol 70:498–507
188. Parveen S, Murphree RL, Edmiston L, Kaspar CW, Portier KM, Tamplin ML (1997) Association of multiple-antibiotic-resistance profiles with point and nonpoint sources of *Escherichia coli* in Apalachicola Bay. Appl Environ Microbiol 63:2607–2612
189. Parveen S, Hodge NC, Stall RE, Farrah SR, Tamplin ML (2001) Phenotypic and genotypic characterization of human and nonhuman *Escherichia coli*. Water Res 35:379–386
190. Payan A, Ebdon J, Taylor H, Gantzer C, Ottoson J, Papageorgiou GT, Blanch AR, Lucena F, Jofre J, Muniesa M (2005) Method for isolation of *Bacteroides* bacteriophage host strains suitable for tracking sources of fecal pollution in water. Appl Environ Microbiol 71:5659–5662
191. Payment P, Franco E (1993) *Clostridium perfringens* and somatic coliphages as indicators of the efficiency of drinking water treatment for viruses and protozoan cysts. Appl Environ Microbiol 59:2418–2424
192. Pianetti A, Sabatini L, Bruscolini F, Chiaverini F, Cecchetti G (2004) Faecal contamination indicators, salmonella, vibrio and aeromonas in water used for the irrigation of agricultural products. Epidemiol Infect 132:231–238

193. Pina S, Puig M, Lucena F, Jofre J, Girones R (1998) Viral pollution in the environment and in shellfish: human adenovirus detection by PCR as an index of human viruses. Appl Environ Microbiol 64:3376–3382

194. Pinto R, Abad F, Gajardo R, Bosch A (1996) Detection of infectious astroviruses in water. Appl Environ Microbiol 62:1811–1813

195. Pruss A (1998) Review of epidemiological studies on health effects from exposure to recreational water. Int J Epidemiol 27:1–9

196. Puig A, Queralt N, Jofre J, Araujo R (1999) Diversity of *Bacteroides fragilis* strains in their capacity to recover phages from human and animal wastes and from fecally polluted wastewater. Appl Environ Microbiol 65:1772–1776

197. Rabinovici SJM, Bernknopf RL, Wein AM, Coursey DL, Whitman RL (2004) Economic and health risk trade-offs of swim closures at a Lake Michigan beach. Environ Sci Technol 38:2737–2745

198. Rajal VB, McSwain BS, Thompson DE, Leutenegger CM, Kildare BJ, Wuertz S (2007) Validation of hollow fiber ultrafiltration and real-time PCR using bacteriophage PP7 as surrogate for the quantification of viruses from water samples. Water Res 41:1411–1422

199. Rajal VB, McSwain BS, Thompson DE, Leutenegger CM, Wuertz S (2007) Molecular quantitative analysis of human viruses in California stormwater. Water Res 41: 4287–4298

200. Ram JL, Thompson B, Turner C, Nechvatal JM, Sheehan H, Bobrin J (2007) Identification of pets and raccoons as sources of bacterial contamination of urban storm sewers using a sequence-based bacterial source tracking method. Water Res 41:3605–3614

201. Resnick IG, Levin MA (1981) Assessment of bifidobacteria as indicators of human fecal pollution. Appl Environ Microbiol 42:433–438

202. Rose JB, Mullinax RL, Singh SN, Yates MV, Gerba CP (1987) Occurrence of rotaviruses and enteroviruses in recreational waters of Oak Creek, Arizona. Water Resour 21:1375–1381

203. Saha ML, Khan MR, Ali M, Hoque S (2009) Bacterial load and chemical pollution level of the River Buriganga, Dhaka, Bangladesh. Bangladesh J Botany 38:87–91

204. Sambrook J, Russell D (2001) Molecular cloning: a labortory manual, 3rd edn. Cold Spring Harbor Laboratory, New York

205. Sandery M, Stinear T, Kaucner C (1996) Detection of pathogenic *Yersinia enterocolitica* in environmental waters by PCR. J Appl Bacteriol 80:327–332

206. Santo Domingo JW, Bambic DG, Edge TA, Wuertz S (2007) Quo vadis source tracking? Towards a strategic framework for environmental monitoring of fecal pollution. Water Res 41:3539–3552

207. Santoro AE, Boehm AB (2007) Frequent occurrence of the human-specific *Bacteroides* fecal marker at an open coast marine beach: relationship to waves, tides and traditional indicators. Environ Microbiol 9:2038–2049

208. Savichtcheva O, Okayama N, Okabe S (2007) Relationships between *Bacteroides* 16 S rRNA genetic markers and presence of bacterial enteric pathogens and conventional fecal indicators. Water Res 41:3615–3628

209. Schriewer A, Miller WA, Byrne BA, Miller MA, Oates S, Conrad PA, Hardin D, Yang HH, Chouicha N, Melli A, Jessup D, Dominik C, Wuertz S (2010) Presence of *Bacteroidales* as a predictor of pathogens in surface waters of the central California coast. Appl Environ Microbiol 76:5802–5814

210. Schulz CJ, Childers GW (2011) Fecal *Bacteroidales* diversity and decay in response to temperature and salinity. Appl Environ Microbiol 77:2563–2572

211. Scott TM, Rose JB, Jenkins TM, Farrah SR, Lukasik J (2002) Microbial source tracking: current methodology and future directions. Appl Environ Microbiol 68:5796–5803

212. Scott TM, Jenkins TM, Lukasik J, Rose JB (2005) Potential use of a host associated molecular marker in *Enterococcus faecium* as an index of human fecal pollution. Environ Sci Technol 39:283–287

213. Sedmak G, Bina D, MacDonald J (2003) Assessment of an enterovirus sewage surveillance system by comparison of clinical isolates with sewage isolates from Milwaukee, Wisconsin, collected August 1994 to December 2002. Appl Environ Microbiol 69:7181–7187

214. Sedmak G, Bina D, Macdonald J, Couillard L (2005) Nine-year study of the occurrence of culturable viruses in source water for two drinking water treatment plants and the influent and effluent of a wastewater treatment plant in Milwaukee, Wisconsin (August 1994 through July 2003). Appl Environ Microbiol 71:1042–1050

215. Semel JD, Trenholme G (1990) *Aeromonas hydrophila* water-associated traumatic wound infections: a review. J Trauma 30:324–327

216. Sercu B, Van De Werfhorst LC, Murray J, Holden PA (2009) Storm drains are sources of human fecal pollution during dry weather in three urban Southern California watersheds. Environ Sci Technol 43:293–298

217. Seurinck S, Verstraete W, Siciliano SD (2003) Use of 16 S-23S rRNA intergenic spacer region PCR and repetitive extragenic palindromic PCR analyses of *Escherichia coli* isolates to identify nonpoint fecal sources. Appl Environ Microbiol 69:4942–4950

218. Seurinck S, Defoirdt T, Verstraete W, Siciliano SD (2005) Detection and quantification of the human-specific HF183 *Bacteroides* 16 S rRNA genetic marker with real-time PCR for assessment of human faecal pollution in freshwater. Environ Microbiol 7:249–259

219. Sghir A, Gramet G, Suau A, Rochet V, Pochart P, Dore J (2000) Quantification of bacterial groups within human fecal flora by oligonucleotide probe hybridization. Appl Environ Microbiol 66:2263–2266

220. Shanks OC, Santo Domingo JW, Lamendella R, Kelty CA, Graham JE (2006) Competitive metagenomic DNA hybridization identifies host-specific microbial genetic markers in cow fecal samples. Appl Environ Microbiol 72:4054–4060

221. Shanks OC, Domingo JW, Lu J, Kelty CA, Graham JE (2007) Identification of bacterial DNA markers for the detection of human fecal pollution in water. Appl Environ Microbiol 73:2416–2422
222. Shanks OC, Atikovic E, Blackwood AD, Lu J, Noble RT, Domingo JS, Seifring S, Sivaganesan M, Haugland RA (2008) Quantitative PCR for detection and enumeration of genetic markers of bovine fecal pollution. Appl Environ Microbiol 74:745–752
223. Shanks OC, Kelty CA, Sivaganesan M, Varma M, Haugland RA (2009) Quantitative PCR for genetic markers of human fecal pollution. Appl Environ Microbiol 75:5507–5513
224. Shanks OC, White K, Kelty CA, Hayes S, Sivaganesan M, Jenkins M, Varma M, Haugland RA (2010) Performance assessment PCR-based assays targeting bacteroidales genetic markers of bovine fecal pollution. Appl Environ Microbiol 76:1359–1366
225. Shanks OC, White K, Kelty CA, Sivaganesan M, Blannon J, Meckes M, Varma M, Haugland RA (2010) Performance of PCR-based assays targeting Bacteroidales genetic markers of human fecal pollution in sewage and fecal samples. Environ Sci Technol 44:6281–6288
226. Shanks OC, Kelty CA, Archibeque S, Jenkins M, Newton RJ, McLellan SL, Huse SM, Sogin ML (2011) Community structure of cattle fecal bacteria from different animal feeding operations. Appl Environ Microbiol 77:2992–3001
227. Shibata T, Solo-Gabriele HM, Fleming LE, Elmir S (2004) Monitoring marine recreational water quality using multiple microbial indicators in an urban tropical environment. Water Res 38:3119–3131
228. Silkie SS, Nelson KL (2009) Concentrations of host-specific and generic fecal markers measured by quantitative PCR in raw sewage and fresh animal feces. Water Res 43: 4860–4871
229. Sinton LW, Hall CH, Lynch PA, Davies-Colley RJ (2002) Sunlight inactivation of fecal indicator bacteria and bacteriophages from waste stabilization pond effluent in fresh and saline waters. Appl Environ Microbiol 68:1122–1131
230. Sivaganesan M, Seifring S, Varma M, Haugland RA, Shanks OC (2008) A Bayesian method for calculating real-time quantitative PCR calibration curves using absolute plasmid DNA standards. BMC Bioinformatics 9:120
231. Sivaganesan M, Haugland RA, Chern EC, Shanks OC (2010) Improved strategies and optimization of calibration models for real-time PCR absolute quantification. Water Res 44:4726–4735
232. Soller JA, Bartrand T, Ashbolt NJ, Ravenscroft J, Wade TJ (2010) Estimating the primary etiologic agents in recreational freshwaters impacted by human sources of faecal contamination. Water Res 44:4736–4747
233. Soller JA, Schoen ME, Bartrand T, Ravenscroft JE, Ashbolt NJ (2010) Estimated human health risks from exposure to recreational waters impacted by human and non-human sources of faecal contamination. Water Res 44: 4674–4691
234. Steets BM, Holden PA (2003) A mechanistic model of runoff-associated fecal coliform fate and transport through a coastal lagoon. Water Res 37:589–608
235. Stewart JR, Ellender RD, Gooch JA, Jiang S, Myoda SP, Weisberg SB (2003) Recommendations for microbial source tracking: lessons from a methods comparison study. J Water Health 1:225–231
236. Stewart JR, Gast RJ, Fujioka RS, Solo-Gabriele HM, Meschke JS, Amaral-Zettler LA, Del Castillo E, Polz MF, Collier TK, Strom MS, Sinigalliano CD, Moeller PD, Holland AF (2008) The coastal environment and human health: microbial indicators, pathogens, sentinels and reservoirs. Environ Health 7(Suppl 2):S3
237. Stoeckel DM, Harwood VJ (2007) Performance, design, and analysis in microbial source tracking studies. Appl Environ Microbiol 73:2405–2415
238. Stoeckel DM, Stelzer EA, Dick LK (2009) Evaluation of two spike-and-recovery controls for assessment of extraction efficiency in microbial source tracking studies. Water Res 43:4820–4827
239. Sullivan D, Brooks P, Tindale N, Chapman S, Ahmed W (2010) Faecal sterols analysis for the identification of human faecal pollution in a non-sewered catchment. Water Sci Technol 61:1355–1361
240. Tartera C, Lucena F, Jofre J (1989) Human origin of *Bacteroides fragilis* bacteriophages present in the environment. Appl Environ Microbiol 55:2696–2701
241. Teng LJ, Hsueh PR, Huang YH, Tsai JC (2004) Identification of *Bacteroides thetaiotaomicron* on the basis of an unexpected specific amplicon of universal 16 S ribosomal DNA PCR. J Clin Microbiol 42:1727–1730
242. Turnbaugh PJ, Hamady M, Yatsunenko T, Cantarel BL, Duncan A, Ley RE, Sogin ML, Jones WJ, Roe BA, Affourtit JP, Egholm M, Henrissat B, Heath AC, Knight R, Gordon JI (2009) A core gut microbiome in obese and lean twins. Nature 457:480–484
243. Ufnar JA, Wang SY, Christiansen JM, Yampara-Iquise H, Carson CA, Ellender RD (2006) Detection of the nifH gene of *Methanobrevibacter smithii*: a potential tool to identify sewage pollution in recreational waters. J Appl Microbiol 101:44–52
244. Ufnar JA, Ufnar DF, Wang SY, Ellender RD (2007) Development of a swine-specific fecal pollution marker based on host differences in methanogen mcrA genes. Appl Environ Microbiol 73:5209–5217
245. Ufnar JA, Wang SY, Ufnar DF, Ellender RD (2007) Methanobrevibacter ruminantium as an indicator of domesticated-ruminant fecal pollution in surface waters. Appl Environ Microbiol 73:7118–7121
246. Unno T, Jang J, Han D, Kim JH, Sadowsky MJ, Kim OS, Chun J, Hur HG (2010) Use of barcoded pyrosequencing and shared OTUs to determine sources of fecal bacteria in watersheds. Environ Sci Technol 44:7777–7782
247. USEPA (2000) Improved enumeration methods for recreational water quality indicators: Enterococci and *Escherichia*

coli. EPA 821/R-97/004. US Environmental Protection Agency Office of Water, Washington, DC

248. USEPA (2002) Method 1603: *Escherichia coli* (*E. coli*) in water by membrane filtration using modified membrane-thermotolerant Escherichia coli agar (modified mTEC) EPA-821-R-02-023. US Environmental Protection Agency Office of Water, Washington, DC
249. USEPA (2002) Method 1604: total coliforms and *Escherichia coli* in water by membrane filtration using a simultaneous detection technique (MI Medium). US Environmental Protection Agency Office of Water, Washington, DC
250. USEPA (2006) Method 1600: Enterococci in Water by Membrane Filtration Using MEMBRANE-Enterococus Indoxyl-B-D-Glucoside agar (mEI) EPA-821-R-06-009. US Environmental Protection Agency Office of Water, Washington, DC
251. USEPA (2007) Critical path science plan for the development of new or revised recreational water quality criteria. 823-R-08-002. US Environmental Protection Agency Office of Water, Washington, DC
252. USEPA (2009) National water quality inventory: report to congress 2004 reporting cycle. EPA 841-R-08-001. US Environmental Protection Agency Office of Water, Washington, DC
253. USEPA (2010) Method A: Enterococci in water by TaqMan® quantitative polymerase chain reaction (qPCR) assay. US Environmental Protection Agency Office of Water, Washington, DC
254. Viau EJ, Goodwin KD, Yamahara KM, Layton BA, Sassoubre LM, Burns S, Tong H-I, Wong SHC, Boehm AB (2011) Human bacterial pathogens and fecal indicators in tropical streams discharging to Hawaiian coastal waters. Water Res 45:3279–3290
255. Volkmann H, Schwartz T, Kirchen S, Stofer C, Obst U (2007) Evaluation of inhibition and cross-reaction effects on real-time PCR applied to the total DNA of wastewater samples for the quantification of bacterial antibiotic resistance genes and taxon-specific targets. Mol Cell Probes 21:125–133
256. Wade TJ, Pai N, Eisenberg JN, Colford JM Jr (2003) Do U.S. Environmental Protection Agency water quality guidelines for recreational waters prevent gastrointestinal illness? A systematic review and meta-analysis. Environ Health Perspect 111:1102–1109
257. Wade TJ, Calderon RL, Sams E, Beach M, Brenner KP, Williams AH, Dufour AP (2006) Rapidly measured indicators of recreational water quality are predictive of swimming-associated gastrointestinal illness. Environ Health Perspect 114:24–28
258. Wade TJ, Calderon RL, Brenner KP, Sams E, Beach M, Haugland R, Wymer L, Dufour AP (2008) High sensitivity of children to swimming-associated gastrointestinal illness: results using a rapid assay of recreational water quality. Epidemiol 19:375–383
259. Wade TJ, Sams E, Brenner KP, Haugland R, Chern E, Beach M, Wymer L, Rankin CC, Love D, Li Q, Noble R, Dufour AP (2010) Rapidly measured indicators of recreational water quality and swimming-associated illness at marine beaches: a prospective cohort study. Environ Health 9:66
260. Walters SP, Gannon VPJ, Field KG (2007) Detection of *Bacteroidales* fecal indicators and the zoonotic pathogens *E. coli* O157:H7, *Salmonella*, and *Campylobacter* in river water. Environ Sci Technol 41:1856–1862
261. Walters SP, Yamahara KM, Boehm AB (2009) Persistence of nucleic acid markers of health-relevant organisms in seawater microcosms: implications for their use in assessing risk in recreational waters. Water Res 43:4929–4939
262. Walters SP, Thebo AL, Boehm AB (2011) Impact of urbanization and agriculture on the occurrence of bacterial pathogens and stx genes in coastal waterbodies of central California. Water Res 45:1752–1762
263. Westrell T, Teunis P, van den Berg H, Lodder W, Ketelaars H, Stenstrom TA, de Roda Husman AM (2006) Short- and long-term variations of norovirus concentrations in the Meuse river during a 2-year study period. Water Res 40: 2613–2620
264. Wetz JJ, Blackwood AD, Fries JS, Williams ZF, Noble RT (2008) Trends in total *Vibrio* spp. and *Vibrio vulnificus* concentrations in the eutrophic Neuse River Estuary, North Carolina, during storm events. Aquat Microb Ecol 53:141–149
265. Whitman RL, Nevers MB (2003) Foreshore sand as a source of *Escherichia coli* in nearshore water of a Lake Michigan beach. Appl Environ Microbiol 69:5555–5562
266. Whitman RL, Nevers MB, Korinek GC, Byappanahalli MN (2004) Solar and temporal effects on *Escherichia coli* concentration at a Lake Michigan swimming beach. Appl Environ Microbiol 70:4276–4285
267. Whitman RL, Ge Z, Nevers MB, Boehm AB, Chern EC, Haugland RA, Lukasik AM, Molina M, Przybyla-Kelly K, Shively DA, White EM, Zepp RG, Byappanahalli MN (2010) Relationship and variation of qPCR and culturable Enterococci estimates in ambient surface waters are predictable. Environ Sci Technol 44:5049–5054
268. Wiedenmann A, Kruger P, Dietz K, Lopez-Pila JM, Szewzyk R, Botzenhart K (2006) A randomized controlled trial assessing infectious disease risks from bathing in fresh recreational waters in relation to the concentration of *Escherichia coli*, intestinal enterococci, *Clostridium perfringens*, and somatic coliphages. Environ Health Perspect 114:228–236
269. Wiggins BA (1996) Discriminant analysis of antibiotic resistance patterns in fecal streptococci, a method to differentiate human and animal sources of fecal pollution in natural waters. Appl Environ Microbiol 62:3997–4002
270. Wilkes G, Edge T, Gannon V, Jokinen C, Lyautey E, Medeiros D, Neumann N, Ruecker N, Topp E, Lapen DR (2009) Seasonal relationships among indicator bacteria, pathogenic bacteria, *Cryptosporidium* oocysts, *Giardia* cysts, and hydrological indices for surface waters within an agricultural landscape. Water Res 43:2209–2223
271. Wong M, Kumar L, Jenkins TM, Xagoraraki I, Phanikumar MS, Rose JB (2009) Evaluation of public health risks at recreational beaches in Lake Michigan via detection of enteric viruses

and a human-specific bacteriological marker. Water Res 43:1137–1149

272. Wu CH, Sercu B, Van de Werfhorst LC, Wong J, DeSantis TZ, Brodie EL, Hazen TC, Holden PA, Andersen GL (2010) Characterization of coastal urban watershed bacterial communities leads to alternative community-based indicators. PLoS One 5: e11285
273. Wyn-Jones AP, Carducci A, Cook N, D'Agostino M, Divizia M, Fleischer J, Gantzer C, Gawler A, Girones R, Höller C, Husman AM, Kay D, Kozyra I, López-Pila J, Muscillo M, Nascimento MS, Papageorgiou G, Rutjes S, Sellwood J, Szewzyk R, Wyer M (2011) Surveillance of adenoviruses and noroviruses in European recreational waters. Water Res 45:1025–1038
274. Xu J, Bjursell MK, Himrod J, Deng S, Carmichael LK, Chiang HC, Hooper LV, Gordon JI (2003) A genomic view of the human-*Bacteroides thetaiotaomicron* symbiosis. Science 299:2074–2076
275. Yamahara KM, Layton BA, Santoro AE, Boehm AB (2007) Beach sands along the California coast are diffuse sources of fecal bacteria to coastal waters. Environ Sci Technol 41:4515–4521
276. Yamahara KM, Walters SP, Boehm AB (2009) Growth of enterococci in unaltered, unseeded beach sands subjected to tidal wetting. Appl Environ Microbiol 75:1517–1524
277. Yampara-Iquise H, Zheng G, Jones JE, Carson CA (2008) Use of a *Bacteroides thetaiotaomicron*-specific alpha-1-6, mannanase quantitative PCR to detect human faecal pollution in water. J Appl Microbiol 105:1686–1693
278. Young TA, Heidler J, Matos-Perez CR, Sapkota A, Toler T, Gibson KE, Schwab KJ, Halden RU (2008) Ab initio and in situ comparison of caffeine, triclosan, and triclocarban as indicators of sewage-derived microbes in surface waters. Environ Sci Technol 42:3335–3340
279. Zheng G, Yampara-Iquise H, Jones JE, Carson CA (2009) Development of *Faecalibacterium* 16 S rRNA gene marker for identification of human faeces. J Appl Microbiol 106:634–641

Marine and Hydrokinetic Energy Environmental Challenges

Andrea E. Copping
Marine Sciences Laboratory, Pacific Northwest
National Laboratory, Seattle, WA, USA

Article Outline

Glossary

Energy from ocean currents Energy harvested from the continuous and predictable flow of ocean currents near landforms.

Marine and hydrokinetic energy Energy that can be harvested from moving water, specifically from ocean waves, tides and currents, and river flow. Abbreviated as MHK, other ocean-based energy resources are usually included in this definition, including ocean thermal energy conversion and energy derived from osmotic gradients.

Ocean thermal energy conversion Energy harvested from the heat differential from cold deep water to warmer surface waters. Abbreviated as OTEC, this energy capture as presently conceived is viable only when a temperature difference of 20°C or more exists from surface to deep ocean water layers.

Osmotic energy Energy harvested from the gradient of salt content from saline ocean water to freshwater. It is generally thought to be viable only at river mouths that empty directly into the sea.

Receptors Terminology for portions of the aquatic ecosystem that are likely to be affected by MHK devices, generally consisting of aquatic animals, habitats, and ecosystems.

River energy Energy derived from the unidirectional flow of water in natural and managed rivers.

Stressors Terminology for parts of MHK systems that may cause stress or harm to aquatic receptors, including the MHK devices, mooring lines, anchors, floats, power cables, and other equipment that may compromise aquatic receptors during installation, operation, maintenance, and decommissioning of MHK developments.

Tidal energy Energy that can be harvested from the daily tidal prism in estuaries, coastal embayments, and other constrictions in ocean areas.

Wave Energy Energy that can be harvested from the vertical motion of waves in coastal areas and/or in open ocean.

Definition of the Subject and Its Importance

The marine and hydrokinetic (MHK) energy industry in the USA and abroad is moving towards commercial-scale development of devices in the marine environment and major rivers in order to generate carbon-free energy in proximity to major load centers along the coastlines. Laws, regulations, and public sentiment require that these deployments and operations provide minimal risk to the marine and riverine environments, the animals that live in them, and to existing human benefits derived from these systems. Effective siting and permitting of MHK devices can only be done successfully by understanding and measuring the environmental effects of MHK devices on the aquatic animals, habitats, and ecosystems of coastal and riverine areas where cost-effective energy generation is possible.

Introduction

As more ocean energy devices are tested, and plans for commercial-scale developments progress, challenges and potential obstacles presented by environmental siting and permitting must be addressed [1–4]. Environmental laws and regulations require that living organisms and systems be protected [5]. US laws and regulations are particularly stringent and require considerable advance understanding of potential effects before devices are deployed [6]. Oceanographic research can provide insight into the action of tides, waves, ocean currents, and other physical and chemical processes in the oceans. Equivalent information about river systems is also available. However, assessments of biological resources in these areas are generally more limited, particularly in the oceans [7]. And very little useful information is available on potential interactions of marine and hydrokinetic devices on marine and freshwater receptors – the animals, habitats, and ecosystems – in areas of proposed marine and hydrokinetic (MHK) energy development [8, 9]. The scientific literature is very light in this area, with most contributions from government reports, gray (non-peer-reviewed) literature, and industry publications. The very high level of uncertainly surrounding potential adverse encounters between MHK systems and vulnerable receptors drives concerns among regulators and stakeholders, potentially slowing development of the MHK industry [10].

This entry will address the major challenges of identifying, measuring, and setting priorities for environmental effects of MHK devices and arrays. First, the various types of MHK devices and the resources they are designed to harvest will be presented. Next, the challenges are introduced as they apply to each of the specific technology types. In the next section, additional challenges to MHK development are briefly presented. Methods for addressing environmental effects of MHK development make up the next section, including a risk-informed approach under development in a US Department of Energy National Laboratory. Mitigation strategies will be discussed. And finally, a path forward to understand and evaluate the challenges of environmental effects of MHK development will be discussed.

Evaluating Environmental Effects of MHK Development

The challenges of predicting the effects and impacts of MHK energy devices on aquatic receptors are hindered by several factors: (1) the technologies are new, and few appropriate analogues exist that will inform potential interactions with living systems; (2) MHK energy development in the ocean is a new use of ocean space, potentially in conflict with other existing uses; (3) MHK device deployment is often planned for areas of fast currents and large waves where there is little oceanographic knowledge; (4) biological resources that are harvested are better known than those which are not directly consumed; however, in many cases, the resources at risk from MHK are not harvested; and (5) MHK energy devices will often be deployed in areas already affected by anthropogenic stressors so that the effect of these devices may be difficult to distinguish from the cumulative impacts of existing industrial activities. Because other industries face similar in-water challenges, environmental effects associated with the installation, decommissioning, and maintenance phases of MHK deployment can be predicted reasonably well. This experience provides a record of interactions with aquatic receptors and suggests appropriate mitigation strategies to minimize risk to the marine and riverine environment [11]. However, effects of the operational phase of MHK deployments are not easily compared to analogues.

MHK devices are tailored for the energy source they will harvest. Although certain hybrid technologies are under development, the major MHK devices are those aimed at harvesting energy from tides, waves, river flow, ocean currents, temperature differential, and salinity differential.

Tides – Unlike earlier attempts to create tidal barrages or tidal fences, most tidal MHK devices are turbines, consisting of a rotor with multiple blades that rotate with the flow of ebb and flood tides. A limited number of bays, estuaries, and passes have good tidal energy potential. Most tidal devices are deployed near the seabed.

Waves – Wave energy converters take many forms and configurations, with most encompassing a surface buoy or float and mooring lines through the water column to anchors in the seabed. This technology is potentially viable wherever waves occur, along coastlines and (for specialized technologies) the open ocean.

River Flow – Most devices designed to capture energy from river flow resemble bladed tidal turbines but are unidirectional and generally smaller than tidal turbines.

Ocean Currents – The ability to harvest energy from constant ocean currents will be accomplished with very large rotating turbines. Ocean currents that are sufficiently constant in location and power density lie along the western edges of ocean basins. Within US waters, the most viable candidate for ocean current harvest is the Gulf Stream along the Florida coastline.

Temperature Gradients – Harvest of energy from temperature gradients in the ocean requires a substantial difference in temperature between warm surface and cool water at depth, as is found in tropical waters. Within US waters, the most viable locations are the Hawaiian Islands and Florida, as well as tropical US territories. Temperature gradients can be harvested by large-scale floating OTEC (ocean thermal energy conversion) plants, equipped with deep (1,000 m or more) cold water uptake pipes and shallower warm water uptake and return pipes.

Osmotic Gradients – The difference in salinity found at river mouths entering coastal seas allows for the harvest of energy from osmotic pressure difference. Very large lengths of fine osmotic exchange tubes are needed for this process; however, only the water intake need be deployed in open water, while the osmotic exchange technology can be packed tightly into buildings on land.

Environmental Challenges Presented by New Technologies

Each MHK technology interacts with the marine and freshwater environment in a different way. Each represents a new series of interactions with the living biota, the habitats, and the ecosystems that will become part of the new power generation system. Some suitable analogues exist to help inform predictions of interactions between the various stressors that make up the MHK installations and the many marine and freshwater receptors, such as offshore oil and gas rigs and navigation buoys creating artificial reefs; interaction of fish in conventional hydropower turbines; and offshore wind turbines changing benthic habitats. However, none of these analogue interactions are exactly indicative of the stressor–receptor interactions expected from arrays of MHK devices in the aquatic environment. Many MHK installations have some attachment to the seabed or riverbed; all place some impediment to water flow in the natural circulation; and all require a power cable to carry electricity to the land-based electrical grid [12].

Separate and distinct stressors will act upon aquatic receptors, depending on the phase of MHK development, from installation, operation, maintenance, and decommissioning. Installation may cause a series of high-intensity interactions that result from placing or anchoring the device to the bottom, including pile driving, pinning tidal turbine foundations, and embedding anchors into bottom sediment. In addition, intensive vessel traffic may disturb marine animals and increase the spill risk of petroleum and other hazardous materials. Decommissioning will also result in increased vessel traffic and may cause benthic habitat destruction from the removal of piles or anchors. Maintenance activities increase vessel traffic, potentially disturbing marine animals and increasing spill risks. Operational effects are generally thought to be less destructive but provide the potential for ongoing interactions between MHK devices and aquatic animals, as well as chronic stressors like acoustic output from turbines and EMF from cables and turbines. Specific environmental interactions for the operational phase of each technology type are summarized here.

Tidal Devices – The most common tidal MHK devices are anchored on or near the seabed with gravity or bottom-piercing mounts. Blades rotate only when the tidal energy reaches the cut-in speed of the device. The benthic habitat onto which the devices are installed is most commonly made up of coarse sediments, as fine sediment particles are carried away by swift currents at the site. Environmental effects of greatest concern for operation of single tidal turbines as well as arrays of devices include: (1) the danger presented by rotating blades to sea life such as marine mammals, fish, and diving birds; (2) acoustic output from rotating turbine blades interrupting marine mammal communication and navigation, as well as disturbing other organisms such as fish and sea turtles; (3) electromagnetic output from turbines and power cables, affecting sensitive organisms, particularly elasmobranchs (sharks and rays); and for large arrays of devices, (4) basin-scale effects from changes in water circulation from the presence of the devices as well as energy removal from power take off, resulting in changes in sediment transport and shore forms, degraded water quality, changes in nearshore habitat quality, and alternation of the base of the marine food web [4, 13, 14]. Maintenance operations generally increase vessel traffic in the area of tidal deployments, potentially disturbing animals and increasing spill risk.

Wave Energy Converters – Wave energy converters (WECs) are most commonly anchored to the seabed with buoys or other structures floating at the surface, attached to the anchors with mooring lines or cables. Environmental effects of greatest concern for operating single WECs as well as arrays of devices include: (1) entanglement of migratory animals such as whales, sea turtles, and diving birds in mooring lines; (2) changes in benthic habitat due to anchors; (3) electromagnetic output from power cables, affecting sensitive organisms, particularly elasmobranchs (sharks and rays) and sea turtles; and for large arrays, (4) changes in benthic habitat and shore forms due to changes in sediment transport, caused by alterations in water circulation around the surface floats [1, 15–17].

River Turbines – Devices planned for river deployment are most commonly turbines similar in design to those used for tidal energy generation, although smaller in size. These devices can be deployed, generally in groups, off bridge abutments, from floating structures, or anchored on foundations above the river bottom. Environmental effects of greatest concern for operation of river turbines include: (1) danger to migratory fish and diving birds from rotating blades; and (2) changes in benthic habitats due to changes in sediment transport, caused by alternations in water circulation around the devices [18].

Ocean Currents – No devices have been tested worldwide; however, these devices are likely to resemble very large tidal turbines and will be deployed at mid-depth in persistent ocean currents, generally at depths of 50–200 m. Surface or subsurface flotation will be needed to keep the turbines at the appropriate depths. Environmental effects of greatest concern for ocean current devices include: (1) entanglement of migratory species particularly whales, dolphins, sea turtles, and large fish in mooring lines, above and below the turbines, as well as avoidance of surface or subsurface floats; (2) changes in benthic habitats due to anchors; and (3) electromagnetic output from turbines and power cables, affecting sensitive organisms, particularly elasmobranchs (sharks and rays) and sea turtles [19].

Temperature Gradients – Pilot-scale OTEC plants were developed several decades ago, and at least one still is in operation in Kona, Hawaii. Full-scale OTEC plants are envisioned as large floating installations anchored to the seafloor in deep water, generally 1,000 m or more. Environmental effects of greatest concern include: (1) changes in the pelagic environment and marine food web due to removal and exchange of warm surface water and cold water from depth; (2) entanglement of migratory species such as marine mammals, fish, sea turtles, and diving birds in mooring lines that reach from the surface installation to the anchors; (3) changes in deep benthic habitats from anchors; and (4) obstructions to migratory pathways due to the large size of the installations [20].

Salinity Gradients – Small-scale osmotic exchange facilities are in operation in Sweden [21] at the mouths of moderate-sized rivers where they enter the sea. To date, the osmotic exchange installations have been housed in buildings onshore, near the river mouths. Environmental effects are likely to be associated with the withdrawal of freshwater from rivers and nearshore areas and the return of water of altered salinity to these bodies of water.

Other Challenges to MHK Energy Development

In addition to direct environmental risks from MHK energy development, there are other challenges that are inextricably associated with the interaction of MHK devices in the oceans and rivers. These challenges include: the interaction of MHK energy development with existing ocean uses; deployment in high-energy waters; limitations of biological assessments in the target development areas; and cumulative environmental effects of MHK development with other anthropogenic and climate stressors.

Ocean Uses – The coastal and riverine areas that are optimal for energy generation, proximity to a grid connection, and the appropriate distance to a port for construction and maintenance may already be in use for other purposes [22]. Coastal oceans are highly prized for commercial and recreational fishing, navigation, shipping and transportation, national security, boating and surfing, marine conservation, waste disposal, and other uses. As a new industry entering the ocean space, the MHK industry faces the challenge of finding the appropriate space needed for commercial development, within the existing spatial layout and potentially in competition with current users [23].

Energetic Marine and Riverine Locations – The estuarine and nearshore ocean areas where energy is sufficient to generate power are often the areas where the natural resources and physical interactions are least well documented [24]. Oceanographers do not routinely place instruments and deploy gear in areas where the tidal currents and waves are likely to damage or destroy expensive equipment, nor do vessels easily hold station for lengthy deployments or measurements in these areas. For these reasons, our knowledge of the physics, chemistry, and biology of potential MHK sites is often incomplete or lacking altogether.

Biological Assessments – Many of the greatest environmental concerns over the development of MHK energy generation surround interactions with marine and freshwater animals [9]. Robust biological assessments are needed to determine the status of populations prior to MHK deployment and to provide a baseline against which post-installation changes can be evaluated [25]. Populations of commercially and recreationally important species (generally fish, shellfish, and whales) are surveyed with some regularity in some coastal waters and rivers, while bird and other wildlife populations are surveyed episodically by government agencies or nongovernmental organizations. However, in waters preferred for MHK development, many of the populations of concern and their habitats are not well documented.

Cumulative Effects – Coastal areas, estuaries, nearshore marine waters, and major rivers are not made up of untouched wilderness; many commercial, recreational, and industrial activities already use the important ecosystem services supplied by these waters. Sorting out the incremental effects that MHK developments may have on the resilience of aquatic animals, habitats, and ecosystems from the cumulative impacts of other anthropogenic activities, even if robust biological and ecosystems assessment data were available, will prove very challenging [19].

Addressing Environmental Effects

The large number of potential interactions between stressors (those parts of MHK systems that may cause stress on aquatic receptors, including the MHK devices, anchors and mooring lines, cables, and flotation) and receptors (aquatic animals, habitats, ecosystems, or the physical and chemical processes that affect them) requires that one determines which stressor/receptor interactions are most significant, which are likely to cause harm to aquatic receptors and for which appropriate mitigation measures can be developed. MHK development will succeed as a viable renewable energy source in an accelerated fashion if collectively a multipronged approach is pursued, which satisfies the need to be protective of marine and freshwater receptors, while allowing adequate testing of devices, mooring lines and anchors, and electrical systems. The conundrum is the dearth of information on stressor/receptor interactions that makes it difficult to ensure safety for aquatic animals, habitats, and ecosystems, while gaining the performance testing data needed to create viable commercial MHK systems.

The scientific community engaged in evaluating environmental effects of MHK systems appears to be coalescing around two approaches: (1) deploying small numbers of devices as rapidly and safely as possible to gather monitoring data about stressor/receptor interactions and (2) examining each likely stressor/receptor interaction for each class of MHK device (i.e., tidal, wave, river, OTEC) and estimating the relative risk of each receptor encounter.

One risk-based approach is being pursued at Pacific Northwest National Laboratory, in collaboration with other US Department of Energy National Laboratories, university partners, with the cooperation of MHK project developers, regulatory agencies, and stakeholders [26]. Risk is being defined as the potential deleterious consequences of stressor/receptor interactions, as a function of consequence (the seriousness of the effect) and a factor of probability (chance that the encounter will occur) [27, 28]. The risk-based approach, dubbed ERES (environmental risk evaluation system), examines every risk-relevant stressor/receptor pair from MHK projects that are in advanced stages of development, using expert opinion to rank the relative consequences among the pairs, in lieu of field and laboratory data. The semiformal gathering of expert opinion uses a series of biophysical risk factors to rank consequence; regulatory requirements are also taken into consideration to determine the importance of each stressor/receptor pair. As experimental monitoring and modeling data become available, the ranking of these consequences will be re-assessed using these data, replacing the expert opinion with results based on empirical studies. The continuing challenge of the system is that there are insufficient data to carry out probability modeling to create a true risk ranking. The output of an initial ERES screening case for a tidal energy project is shown, for illustration purposes only, in Table 1. Consequence rankings within a tier are considered tied ranks.

Following the expert opinion-based ranking of stressor/receptor pairs for consequences, probability analysis will determine whether some of the highest ranked stressor/receptor pairs represent low probability of encounter and, therefore, a reduced risk despite the potentially high consequences; additionally, further elucidation of the tied ranks will be possible with probability analysis, once sufficient data exist. As more data become available, it will become clear which of the

Marine and Hydrokinetic Energy Environmental Challenges. Table 1 Top tiers of risk based on environmental consequence for stressor/receptor pairs for a tidal case, using the environmental risk evaluation system (ERES). Each receptor group is listed, with an example from Pacific NORTHWEST location in the USA, in brackets. T&E – threatened and endangered animal, listed under the Endangered Species Act (After [26], 53 pp)

Tidal stressor	Tidal receptor
First tier	
Accident/disaster (oil spills)	T&E cetacean (killer whale)
Accident/disaster (oil spills)	T&E pinniped (Steller sea lion)
Physical presence (dynamic)	T&E cetacean (killer whale)
Physical presence (dynamic)	T&E pinniped (Steller sea lion)
Physical presence (dynamic)	T&E bird (marbled murrelet)
Second tier	
Noise	T&E cetacean (killer whale)
Noise	T&E pinniped (Steller sea lion)
Accident/disaster (oil spill)	T&E bird (marbled murrelet)
Leaching of toxic chemicals	T&E bird (marbled murrelet)
Noise	T&E bird (marbled murrelet)
Third tier	
EMF	T&E bird (marbled murrelet)
Physical presence (static)	T&E bird (marbled murrelet)
Fourth tier	
Physical presence (static)	T&E cetacean (killer whale)
Physical presence (static)	T&E pinniped (Steller sea lion)
Fifth tier	
EMF	T&E cetacean (killer whale)
EMF	T&E pinniped (Steller sea lion)
Leaching of toxic chemicals	T&E cetacean (killer whale)
Leaching of toxic chemicals	T&E pinniped (Steller sea lion)

highest ranked stressor/receptor interactions are due to great uncertainty, and which truly constitute risk of adverse impacts to aquatic receptors. Stressor/receptor pairs that rank high due to uncertainty will benefit from additional research and monitoring effort, perhaps allowing some to become less highly ranked. As highly ranked stressor/receptor pairs move to lower tiers with additional information, those pairs that remain highly ranked will be candidates for mitigation. In this iterative manner, ERES will assist MHK developers, regulators, and stakeholders to focus on mitigating the most severe risks and allow renewable, carbon-free MHK energy generation to accelerate.

Regulatory Drivers – In our society, the relative level of concern over environmental effects is determined through the regulatory process. Laws and regulations are enacted to protect natural resources and human health and create de facto value systems for society's use of resources. In order to regulate a new industry, with technologies that have no close analogues, the existing laws and regulations will need to be adapted to cover MHK development. These regulatory drivers often are not well aligned with the realities of new technologies such as MHK and may trigger concerns over risks to the marine and freshwater environment that later are shown to be relatively unimportant. As more experience with MHK devices in the water and more information about environmental effects become available, more impact-appropriate approaches to regulating the industry are likely to emerge, as has happened in many other resource-based industries. Variations on existing regulatory processes that have been implemented in other countries include provisions for early demonstration or small-scale pilot projects that are believed to have minimal impact on the environment; the rationale behind these early deployments is to gather experience and information to better inform commercial-scale environmental risks from MHK devices. However, as the MHK industry emerges, in most countries including the USA, application of existing laws and regulatory pathways will continue to be the major construct for advancing towards commercial development.

Mitigation Strategies

As environmental risks to aquatic animals, habitats, or ecosystems are identified through directed research and monitoring of in-water MHK sites, strategies to mitigate the impacts must be devised and implemented to reduce risks and protect vulnerable living systems. Although little formal investigation of mitigation strategies has been carried out for MHK deployments and operations, they appear to fall into four categories of mitigation, including siting, engineering design, modification of animal behavior, and exclusion of receptors from interaction. Each strategy will be briefly discussed.

Siting as Mitigation Strategy – At a gross level, MHK devices and farms must be located in areas where there is sufficient power resource to allow for generation of power, at whatever level is desired (small distributed systems, utility-scale production, etc.). Siting also requires that attention be paid to distance from shore and capable ports and harbors such that power cables can be laid to shore in proximity to transmission interconnections and to allow for cost-effective and safe installation and maintenance. However, within these constraints, there are generally options for micro-siting and displacement from original plans to lower risk to aquatic animals, habitats, or the ecosystem. For example, if a wave installation is planned for an area where large numbers of seabirds congregate, moving the installation a few kilometers offshore could lower the risk to the birds while causing little disruption to the power-generating potential. Similarly, during preliminary planning, the relocation of the foundation of a tidal turbine or wave anchor from a known crab nursery ground would greatly reduce the environmental risk at little or no increased cost of installation or power production.

Engineering Design Feedback as Mitigation – As environmental studies and risk evaluation of MHK projects proceed in parallel to engineering of deployment, anchoring, and operational systems, there are opportunities for feedback from the environmental risk analysis to engineering teams to reduce stress to the environment with engineering and design solutions. For example, if the leading edge of a river or tidal turbine blade was shown (or even strongly suspected) of causing a threat to a migrating endangered fish species, that knowledge would allow the engineering team to redesign the blade shape or operation to lower that risk. Another example includes the tautness of mooring lines for wave buoys could be shown to affect the frequency with which marine mammals become entangled, allowing the engineers designing the lines, materials, and operational status to make changes that will protect the animals. And yet another example might be the discovery that EMF from power cables is causing disruptions to feeding and movement of fish such as sharks in marine waters and sturgeon in freshwater; engineering solutions such as cable burial, switching from AC transmission to DC, or specialized shielding, could decrease or eliminate the risk to fish.

Behavior Modification of Threatened Animals – Risk evaluation will help to pinpoint specific aspects of MHK devices and systems that are particularly dangerous to aquatic animals. This knowledge can be used to deter at risk animals from approaching close enough to the devices or systems for that encounter to take place. For example, acoustic deterrents are under investigation on lines of wave buoys to deter migrating whales traveling parallel to the coast, with the desired outcome to move the whales slightly further offshore. Similarly, seabirds can be encouraged to avoid wave platforms with the use of acoustic or visual deterrents. This form of mitigation holds great promise of reducing targeted risk to aquatic animals; however, there are two barriers to its widespread use: (1) predicting the behavior of animals encountering new stimuli in their habitat is very uncertain, leading to potentially unintended consequences, such as animals being preferentially attracted rather than repelled by warnings, or becoming desensitized over time; and (2) harassment of many aquatic animals, which includes injury, death, or alterations in behavior, is not permitted with few exceptions; harassment applies to most protected species (in the USA under the Endangered Species Act or in the EU under Natura 2000), as well as most or all marine mammals (in the USA under the Marine Mammal Protection Act). Additional research is needed into safe deterrents and behavioral reactions of vulnerable animals; this research generally must be carried out under the auspices of the government agencies responsible for protecting these species.

Exclusion of Aquatic Receptors from MHK Devices – Each of the previously discussed mitigation strategies seeks to decrease the risk to aquatic animals, habitats, and ecosystems from stressors created by MHK devices and systems. However, there is an increasing focus from regulatory bodies and stakeholders on taking more proactive steps to ensure the safety of these receptors. Physical barriers and operational changes to prevent the opportunity of encounter are most commonly considered, although other options may be considered as well. Most prominently considered are processes that recognize the threat to an aquatic animal in real time and take action to shutdown a rotating turbine. The pioneering work on the marine current turbine in Strangford Lough in Northern Ireland has involved marine mammal observers, making visual observations during daylight hours to trigger a shutdown of the turbines when endangered pinnipeds approach, and later the use of sonar to provide the approach data to the observer to trigger a shutdown. Another example is the development of the marine animal alert system (MAAS) at PNNL; the MAAS will use passive and active acoustic signals to detect and classify members of an endangered small whale population in Puget Sound and provide a signal to automatically shutdown the nearby tidal turbines when the whales enter a risk envelope developed in partnership and cooperation with the regulatory body. These mitigation strategies, and similar plans to exclude animals from encounters, provide a strong measure of safety for the aquatic receptors but may have sizable deleterious consequences to the ability to develop the MHK industry as a viable renewable energy source.

Conclusions and Future Directions

The MHK industry is emerging in Europe, Asia, North America, and Brazil, based on the extremely large potential for power generation from waves, tides, river flows, ocean current, and other attributes of the world's water resources [29]. There are many MHK technology designs in various stages of development, with a few ready for commercial-scale deployment while most are still in the testing and development stages. As the formidable hurdles of designing robust systems for long-term deployment in harsh ocean and river conditions proceed, the challenges of protecting the natural environment must be given a prominent place in the development sequence. The laws and regulations governing deployment in the water, bringing power cables to land through the nearshore and intertidal/riparian zone, and the activities surrounding installation, operation, maintenance, and decommissioning of these systems require understanding the baseline conditions where the MHK devices will be placed and the effects and impacts once they are in place. Appropriate mitigation of impacts that cannot be eliminated will become lifelong features of MHK farms. In addition, stakeholders will hold the industry and regulators responsible for adverse outcomes and will press for a high standard of certainty before commercial-scale MHK farms are developed.

Many of the information needs, strategies for collecting cost-effective and accurate monitoring data over long periods of time, and devising effective and efficient mitigation strategies can be assisted with deliberate research programs coordinated among the MHK industry, regulators, funding agencies, and the research community. Early and open coordination of studies, monitoring designs, and mitigation strategies among the interested parties is necessary. The broader the discussions and coordination of these goals and strategies is, and the engagement of all parties, the faster a sustainable and vibrant MHK industry, contributing to the renewable energy portfolio, will become a reality.

Bibliography

1. Boehlert G, McMurray G, Tortorici C (2008) Ecological effects of wave energy development in the Pacific Northwest: a scientific workshop. U.S. Department of Commerce, Seattle, 11–12 Oct 2007
2. Cada G, Ahlgrimm J, Bahleda M, Bigford T, Stavrakas SD, Hall D, Moursund R, Sale M (2007) Potential impacts of hydrokinetic and wave energy conversion technologies on aquatic environments. Fisheries 32:174–181
3. Dadswell MJ, Rulifson RA, Daborn GR (1986) Potential impact of large-scale tidal power developments in the upper Bay of Fundy on fisheries resources of the northwest Atlantic. Fisheries 11:26–35
4. Polagye B, Copping A, Kirkendall K, Boehlert G, Walker S, Weinstein M, Cleve BV (in press) Environmental effects of tidal energy development: a scientific workshop. University of Washington, Seattle, 22–24 Mar 2010
5. Federal Energy Regulatory Commission (2008) White paper on licensing hydrokinetic pilot projects. Federal Energy Regulatory Commission, Washington, DC
6. Michel J, Dunagan H, Boring C, Healy E, Evans W, Dean J, McGillis A, Hain J (2007) Worldwide synthesis and analysis of existing information regarding environmental effects of alternative energy uses on the outer continental shelf. Minerals Management Service, U.S. Department of the Interior, Washington, DC
7. U.S. Commission on Ocean Policy (2004) An ocean blueprint for the 21st century. Final report, Washington, DC
8. Electric Power Research Institute (2008) Prioritized research, development, deployment and demonstration (RDD&D) needs: marine and other hydrokinetic renewable energy. EPRI, Palo Alto
9. U.S. Department of Energy (2009) Report to congress on the potential environmental effects of marine and hydrokinetic energy technologies. Wind and Hydropower Technologies Program, U.S. Department of Energy, Washington DC
10. Copping AE, Geerlofs S (2010) Report on outreach to stakeholders for fiscal year 2009. Pacific Northwest National Laboratory, Seattle
11. Bedard R, Previsic M, Hagerman G, Polagye B, Musial W, Klure J, von Jouanne A, Mathur U, Collar C, Hopper C, Amsden S (2007) North American ocean energy status – Mar 2007. In: European wave and tidal energy conference, Porto, pp 1–8
12. Gill A (2005) Offshore renewable energy: ecological implications of generating electricity in the coastal zone. J Appl Ecol 42:605–615
13. Dacre SL, Bryden IG, Bullen CR (2002) Environmental impacts and constraints of tidal current energy: the Pentland Firth feasibility study. In: Proceedings from MAREC 2002 two day international conference on marine renewable energy, Newcastle
14. El-Geziry TM, Bryden I, Couch S (2009) Environmental impact assessment for tidal energy schemes: an exemplar case study of the Strait of Messina. Proc IMarEST Part A J Mar Eng Technol 13:39–48
15. Bald J, del Campo A, Franco J, Galparsoro I, González M, Liria P, Muxika I, Rubio A, Solaun O, Uriarte A, Comesaña M, Cacabelos A, Fernández R, Méndez G, Prada D, Zubiate L (2010) Protocol to develop an environmental impact study of wave energy converters AZTI-Tecnalia, Herrera Kaia
16. Electrical Power Research Institute (2004) Offshore wave power in the U.S.: Environmental issues. EPRI, Edison
17. Grecian WJ, Inger R, Attrill MJ, Bearhop S, Godley BJ, Witt MJ, Votier SC (2010) Potential impacts of wave-powered marine renewable energy installations on marine birds. IBIS 152: 683–697
18. Cada G, Copping A, Roberts J (in press) The U.S. Department of Energy's efforts to identify and resolve environmental impacts of marine and hydrokinetic energy technologies. Hydrovision
19. Minerals Management Service, Renewable Energy and Alternate Use Program, U.S. Department of the Interior (2006) Technology white paper on ocean current energy potential on the U.S. outer continental shelf. U.S. Department of the Interior, Minerals Management Service, Renewable Energy and Alternate Use Program, Washington, DC
20. Coastal Response Research Center (2010) Ocean thermal energy conversion: assessing potential physical, chemical and biological impacts and risks, Durham
21. Skilhagen SE, Dugstad J, Aaberg R (2008) Osmotic power – power production based on the osmotic pressure difference between waters with varying salt gradients. Desalination 220:476–482
22. Nelson PA, Behrens D, Castle J, Crawford G, Gaddam RN, Hackett SC, Largier J, Lohse DP, Mills KL, Raimondi PT, Robart M,

Sydeman WJ, Thompson SA, Woo S (2008) Developing wave energy in coastal California: potential socio-economic and environmental effects. California Energy Commission, Sacramentro

23. Ehler C, Douvere F (2009) Marine spatial planning: a step-by-step approach toward ecosystem-based management. UNESCO, Paris
24. OEER Association (2008) Fundy tidal energy strategic environmental assessment. Final report, Halifax
25. Shields MA, Dillon LJ, Woolf DK, Ford AT (2009) Strategic priorities for assessing ecological impacts of marine renewable energy devices in the Pentland Firth (Scotland, UK). Mar Policy 33:635–642
26. Copping AE, Van Cleve FB, Anderson RM (2011) Preliminary screening analysis for the environmental risk evaluation system. Evaluating effects of stressors. Environmental effects of marine and hydrokinetic energy. Report to the U.S. department of energy, wind and waterpower program. Pacific Northwest National Laboratory, Seattle, WA, 53 pp
27. US Environmental Protection Agency (1998) Guidelines for ecological risk assessment. US Environmental Protection Agency, Washington DC
28. US Environmental Protection Agency (2003) Framework for cumulative risk assessment. U.S. Environmental Protection Agency, Washington, DC
29. Musial W (2008) Status of wave and tidal power technologies for the United States. National Renewable Energy Laboratory, Golden

Marine Aquaculture in the Mediterranean

Dror L. Angel
Leon Recanati Institute for Maritime Studies,
University of Haifa, Mt Carmel, Haifa, Israel

Article Outline

Glossary

Bioassay (BIOlogical ASSAY) A procedure to test the effect of a substance on living organisms, e.g., the effect of plant nutrients on plant growth rate.

Chemotherapeutants The use of chemicals to treat disease.

Dead zones Coastal areas that undergo seasonal hypoxia (low-oxygen), generally related to eutrophication events, whereafter many of the local (mainly benthic) animals die.

Exotic species An introduced or alien species living outside its natural range, which has been introduced by deliberate or accidental human activity.

FCR (feed conversion ratio) The efficiency at which an animal converts its food into biomass (body mass); FCR = mass of food eaten/increase in biomass.

Immunostimulants Chemicals used to stimulate the immune system by inducing activation or increasing activity of any of its components.

Marine protected areas Areas that restrict human activity (e.g., fishing, boating, coastal development) to protect living, nonliving, cultural, and/or historic resources.

NIMBYism "Not In My Back Yard"-ism; the practice of objecting to a human activity (generally commercial or industrial) that will take place near one's home.

Oligotrophic Waters that have low levels of nutrients and algae, high level of dissolved oxygen, and deep light penetration (i.e., clarity).

Prebiotics Food ingredients (e.g., soluble fiber) that stimulate the growth and/or activity of bacteria in the digestive system which are beneficial to the health of the body.

Probiont Living bacteria added to the environment and feed of reared animals and thought to benefit them by improving intestinal microbial balance, thereby inhibiting pathogenic bacteria.

Protista Unicellular (single-cell) eukaryotic organisms, e.g., foraminifera.

Definition of the Subject

Fisheries and aquaculture play an important role in the economies of many countries; yet this fact is often overlooked as the focus, in many nations, is on provision of food primarily, if not exclusively, from terrestrial agriculture. The value of seafood products as a source of foreign currency is especially important in developing countries and in many cases may exceed the profits from certain agricultural products [1], though this fact also tends to evade common knowledge. The Mediterranean aquaculture sector continues to grow at a rate of close to 9% per year (since 1970) as compared to 3% per year for farmed meat production systems. If the growth of the aquaculture sector can be sustained, it is likely to fulfill the demand for aquatic food supplies by supplying >50% of the total aquatic food consumption within the next 5 years! Therefore, the emphasis here is on the review of the sustainable growth of a commercial activity within an enclosed sea with many conflicting multinational interests. Aquaculture includes the cultivation of finfish, shellfish, crustaceans, and algae; however, this review will focus primarily on Mediterranean finfish farming since many of the sustainability issues revolve around fish farms. There are many different facets (e.g., ecological, social, political, economic) to sustainable commercial activities and this review will touch on several, though not all, of the issues related to aquaculture and its sustainable development in the Mediterranean Sea region.

Introduction

The Mediterranean Sea Environment

Although the term "environment" is often used to mean "ecology," the following description embraces the more holistic meaning, which includes the socioeconomic aspects as well. The Mediterranean is a large, semi-enclosed sea bordered by 22 countries, with two distinct basins divided by a narrow, relatively shallow channel between Sicily in the north and Tunisia in the south. The areal division of the sea between the western and eastern basin is roughly 1/3:2/3. The eastern basin is somewhat more saline than the western basin, especially in the vicinity of the Suez Canal. The Mediterranean Sea has a wide range of seawater temperatures, from as low as 5°C in the Gulf of Trieste in the winter to 31°C off the coast of Libya in the summer [2]. The sea is oligotrophic and phosphorus limited [3] though some limited areas (such as parts of the northern Adriatic) may be eutrophic and it is warmer and more oligotrophic in its southern and eastern areas. Whereas the Mediterranean Sea accounts for only 1% of the world's ocean, it contains 6% of the world's marine species, including >400 endemic species of plants and animals [4]. Despite this impressive biodiversity, biomass is relatively low, mainly due to low primary production.

There are approximately 82 million people in the Mediterranean coastal zone: most in coastal cities and 32% of the population is in North Africa. Levels of development vary widely over the region. Tourism brings >100 million visitors to coastal areas annually, serving as a major source of seasonal population pressure and income and is thus a major competing sector with aquaculture. The Mediterranean Sea is a major shipping route, bridging between Europe and the Middle East and is a base for capture fisheries and mariculture. There are 75 marine protected areas (MPA) in the region, designed to protect unique and threatened resources and habitats such as the seagrass *Posidonia oceanica*, and breeding and nesting sites for endangered species, such as the loggerhead sea turtle (*Caretta caretta*). MPAs were also designated to encourage specific uses, such as sustainable tourism and regenerating fish stocks [5].

A Brief History of Mediterranean Aquaculture

The earliest evidence of aquaculture activity in the Middle East is from the ancient Egyptians. An Egyptian frieze, dated from 2500 B.C., depicted men gathering fish from a pond in what may be the earliest record of such activities in this region [6, 7]. In the sixth and fifth centuries B.C., the Etruscans reared fish in marine farms and the Greeks grew mollusks [8]. Throughout the Roman empire, marine fish (mainly sea bass, sea bream, and mullets) and oysters were reared in special enclosures (e.g., piscines) along the coast [9–11], but this practice seems to have died out with the collapse of the empire and did not appear in the Mediterranean until the middle ages. It is not clear precisely when it began, but there are records of extensive aquaculture in lagoons in Italy, also known as valliculture, starting from around the fifteenth century. Europeans traditionally collected shellfish along the shores, but since the eighteenth century the French oyster industry added a more reliable source – shellfish reared in specialized gear in the intertidal zone. Shellfish aquaculture expanded in the nineteenth century and coastal cultivation spread throughout the Western Mediterranean and the northern Adriatic Sea.

In the second half of the twentieth century, aquaculture developed rapidly, mainly as a result of successful research into the life cycle of the farmed animals (reproduction and larval rearing), as well as physiology, nutrition, and engineering of farming systems [8].

Main Forms of Mariculture (Culture Types and Species) in the Mediterranean

On a global scale, aquaculture production in the Mediterranean Sea is small, but not insignificant – especially with regard to the European demand for fresh seafood. Total aquaculture production in the Mediterranean Sea in 2006 was about 370,000 t [1] with 14% growth from 2000 to 2006, outpacing the growth of capture fisheries. It is noteworthy that the interannual variability in aquaculture production is lower than in capture fisheries (these have reached a plateau in terms of annual harvest), which may be a consideration of prime significance for business and decision-makers concerned with food security, coastal communities, and development.

Within the Mediterranean aquaculture sector, the most striking feature of production is the rate at which finfish have overtaken mussels as the dominant product. In 1990, finfish production accounted for less than 10,000 t as compared to approximately 90,000 t of mussels. In 2003, 180,000 t finfish and 150,000 t mussels were produced (49% and 40% of total production, respectively). Clam and oyster production were only 7% and 2%, respectively, and the remainder of production (~2%) was crustaceans and seaweed. The main cultivated finfish species in the region are gilthead sea bream (*Sparus aurata*), European sea bass (*Dicentrarchus labrax*), and flathead gray mullet (*Mugil cephalus*). Greece, Turkey, Spain, and Italy were the four largest producers of sea bream and bass in 2006, comprising >90% of total Mediterranean production. Sea bream and bass are predominantly reared in net cages in coastal waters, whereas mullets are generally reared in ponds. The major producers of mullets are Egypt and Italy with Egypt generating more than 90% of global mullet production.

A fairly recent development is the farming of bluefin tuna in the Mediterranean, which mainly serves the Japanese sushi market. Tuna farming falls in between the definitions of a standard fishery, which is defined as "capture of wild stock" and aquaculture where fish are both bred and reared in captivity. Because tuna farming is a "postharvest" practice, it is not governed by the regulations of GFCM or ICCAT [12] and as a result there was unregulated growth in this sector, putting heavy pressure on the endangered Mediterranean wild stocks. Concerted efforts are being made to create brood stocks and hatcheries to enable the cultivation of bluefin tuna by the traditional aquaculture methods to release pressure on the endangered Mediterranean wild stocks.

Sustainable Marine Aquaculture in the Mediterranean

One of the features of marine aquaculture in the Mediterranean is that it is developing rapidly in response to a large and ever-growing demand for seafood. This demand was traditionally supplied by fisheries, but the drop in landings in recent decades as a result of overfishing has opened the path for sustainable alternatives to provision of seafood, namely aquaculture. That said, mariculture needs to operate in a manner that will minimize negative impacts on the marine environment, on wild stocks, and on other uses of the seas. Thus, sustainable aquaculture must ensure "*economic viability, social equity and acceptable environmental impacts*" [13].

It is obvious that aquaculture activity must be profitable to succeed, but there are many criteria to profitability and *economic viability* and these may vary considerably in countries that are at different stages of economic development (the process whereby an economic activity develops the technology and experience needed to operate successfully) or that have different interests in mind. In some developing countries, aquaculture may serve as a much needed food and protein source for local consumption, whereas other developing countries may prefer to export their aquaculture production for economic benefit.

Another component of sustainability is *social equity.* Societal equity depends on cultural norms and tendencies of society and varies considerably among the Mediterranean countries. It is probably the most difficult aspect of sustainability to consider because of its intrinsic variability.

Environmental "*acceptability*" is also a difficult issue because of the obvious question: "acceptable by whom?" In order to address this, one needs to consider where the aquaculture activity takes place, who are the stakeholders and how this activity may be conducted in such a manner that it will be acceptable by as many stakeholders as possible. The first aspect of sustainability, discussed below, is the public perception of aquaculture since public opinion may play an important role in the success or failure of the industry. In addition to the various social ramifications, "environmental acceptability" includes the effects of aquaculture on its surroundings and on the ecosystem. The following sections list several of the environmental issues that affect or are affected by Mediterranean aquaculture and a discussion of what is being done about them to enhance the sustainability of this sector.

Public Perception of Aquaculture

The *image* of fish farming varies considerably among different countries and can have a strong effect on the sustainability of the industry. In some northern European countries, the public considers aquaculture in a positive light as a means to enhance food safety and security. In comparison, many southern European countries have a generally negative attitude toward farmed fish as these are considered inferior in taste and health value in comparison to wild-caught ("natural") fish [14, 15]. Numerous negative connotations are associated with marine aquaculture, including: "pollution causing eutrophication," "discharge of antibiotics and harmful chemicals into the environment," "genetic dilution/pollution of wild fish stocks," and "negative visual impact on the coasts."

The public perception is very important for both producers and coastal zone managers since there are many factors that are stacked against the aquaculture sector [16, 17]. These include lack of knowledge on many aspects of the coastal environment, the weakness of a small industry, competition with tourism and other coastal stakeholders, and increasing political power of local environmental lobbies and associations. These lead to non-sustainable situations, including loss of licenses, leases and markets, and reduced diversity in the coastal economy.

The social acceptability of aquaculture was examined at two Greek islands [18] and revealed that residents were more likely to be opposed to aquaculture if they thought that the fish farms would pollute the environment. A study conducted in Israel [19] evaluated public attitudes toward aquaculture and concluded that although most citizens were not terribly well informed in the implications of aquaculture on tourism and environmental issues, the majority are in favor of marine aquaculture. It is noteworthy that this lack of familiarity with aquaculture and aquaculture implications was also observed among the public surveyed in such countries as Scotland [20], Australia [16], and Germany (Schultz, unpublished).

Although the above focuses on the attitudes of the lay public toward aquaculture, it is possible that the opinion of stakeholders is equally (or more) important, despite the fact that the number of stakeholders is usually smaller. Competition over the coastal zone is one of the major sustainability issues that Mediterranean aquaculture faces on a regular and large-scale basis. The competition is especially severe between aquaculture and tourism since the Mediterranean attracts about 30% of the volume of global tourism annually and this is expected to increase over time. There are many examples of such competition, and one of the more recent clashes between the tourism and aquaculture sectors occurred in Turkey in 2008–2009, resulting in a major shift in legislation and in aquaculture lease requirements.

Measures to Improve the Public Attitudes Toward Aquaculture The negative attitudes toward aquaculture are largely a result of ignorance. The media often presents NGO views and opinions in their description of the fish-farming industry, and many of the facts presented are incorrect. The way to correct some of the misconceptions surrounding aquaculture is by preparing a well-planned outreach and educational program geared to reach as many households as possible. There are myths and misconceptions regarding such things as how fish are reared and the densities at which they are stocked, the safety of the feed used, the quality and healthiness of farmed versus wild fish, etc. Preparation of an aquaculture "module" to be taught at schools is an effective way to reach and educate future stakeholders and decision-makers. Another measure that could reduce conflict between aquaculture and other coastal stakeholders is a search for synergies among the stakeholders that would enable multiple use of the coastal zone [21]. Promotion of organic and other types of certification programs to increase public confidence in aquaculture practices and products would also improve public attitude toward this sector.

Benthic Impacts

In the 1990s, the study of the interactions of Mediterranean marine aquaculture with the environment focused on the negative impacts of the industry since most of the early research on salmon farms documented heavy benthic loading, which caused serious damage to underlying seafloor communities and in some cases to the water column as well [22–26]. Benthic organic enrichment that often occurs under intensive finfish farms rapidly leads to hypoxia and anoxia in the sediments. Anoxic sediments support bacterial sulfate reduction, generally leading to an increase in sediment hydrogen sulfide [27]; conditions that are noxious, at best and often lethal to macro- and meiofauna [28]. Although abundances of macrofauna in Mediterranean sediments are considerably lower than the abundances found in temperate regions [29–31], defaunation under fish farms strongly reduces benthic bioturbation (i.e., aeration of the sediments) and leads to accumulation of reduced compounds and organic matter therein. If the farm is situated at a site with limited flushing and circulation, the depth and aerial extent of the impacted sediments may grow with time, creating localized "dead zones." Moreover, when methane accumulates in and bubbles out of anoxic sediments, noxious chemicals such as ammonia and hydrogen sulfide may affect the cultivated fish in the overlying cages.

Because the Mediterranean Sea is largely oligotrophic, and fish farming is generally not practiced at sites with poor flushing, the phenomena described above are not common. At a few sites with limited water circulation, for example, some farms in Croatia and Greece, organic enrichment of the seafloor and local impacts were observed, but these were exceptional and sediment conditions under Mediterranean fish farms are generally less impacted.

At those sites that showed evidence of impacted sediments, the visible effects generally did not extend beyond tens of meters from the edge of the perimeter of the farm [32], though the situation at each farm is different as a result of site-specific currents, depth, bathymetry, etc. The determination of the extent of impacted sediments and benthos (distance from the farm) is subjective and may be strongly affected by the method used. Organic matter determinations, visual inspection, and macrofauna indices are often the methods used to assess the state of the sediments and these clearly show a local effect that diminishes with increasing distance from the point source. However, more sophisticated analyses involving stable isotope signatures of farm effluents indicate that the aquaculture effluents may be detected as far away as 1–2 km from the farms [33–35]. It is very important to qualify the meaning of these measurements because they may be used to make a point about the extent of fish farm effects, but the real issue at hand is the extent of "significant impact." The distribution of small suspended particles over great distances will only constitute a significant impact if the flux of these particles is large and in the case of Mediterranean fish farms, the flux of very small suspended particles is small [36]. Therefore – it is essential to emphasize the difference between qualitative and quantitative effects.

Measures to Reduce Benthic Impacts Despite the fact that benthic loading is generally not a major issue in the Mediterranean, a number of different approaches are employed to increase feeding efficiency and reduce benthic loading. Feeding efficiency is not only an environmental issue, but also a major economic consideration since one of the greatest cost factors in intensive fish farming is the formulated feed. Feeding efficiency includes optimizing the composition of the feed (optimal digestibility) to maximize growth and minimize waste at the lowest possible cost, as well as feed delivery. Considerable efforts are invested by feed companies and fish nutritionists to optimize feed for the various strains of cultivated Mediterranean finfish [37, 38] and during recent years, sea bream and sea bass feed conversion ratios (FCR) have been substantially improved, largely (though not exclusively) due to improved diets and feed delivery. Feed delivery includes the optimal feeding regime whereby feed is provided to the caged fish in suitable portions and at the correct intervals to both maximize growth and health and minimize loss to the surrounding waters. Low-tech feeding involves delivery of pelleted feed to fish either manually by hand, or with the aid of a compressor and regulating the amount according to the response of the fish. High-tech systems include feeding programs that are computerized and customized to each individual cage to optimize delivery of feed to the stock. Another sophistication is the use of submerged Doppler systems (e.g., Doppler Pellet Sensor) that detect when fish stop feeding (increase in the flux of pellets to the bottom of the cages), and send signals to cause the automated feeders to cease feeding (http://www.akvagroup.com/). Many of the above are technologies that were developed outside of the Mediterranean, but as they are also applicable to sea bream and bass production, they are widely used by this sector. One of the more recent developments in Mediterranean aquaculture was the tuna-fattening process, which offered large profits to the farmers. Although it is arguable whether this process should actually be qualified as aquaculture, the environmental ramifications were clear. The penned fish are fed freshly caught or frozen fish rather than pelleted feed and release large amounts of waste (greater than would be released from pelleted food) to the seafloor and have rather high feed conversion ratios (FCRs) 10:1–20:1 as compared to the FCR of sea bream (2:1) or salmon (1:1). Research is currently ongoing to develop artificial diets to create a better FCR for the tuna and to reduce the reliance and fishing pressure on small pelagic fish (e.g., [39]).

Water Quality

The sustainability of any human activity is a function of the nature of the receiving or host environment and in the case of aquaculture this is the basis for estimating the assimilative, holding, or carrying capacity [40]. At a few sites with restricted water exchange, for example, lagoons, there were reports of eutrophication problems [41–43] as the loading of organic and inorganic nutrients clearly exceeded the capacity of the environment to assimilate these [44]. Sites where such self-pollution problems emerge suggest that the preliminary environmental impact assessment and site selection procedures were not carried out properly. In the oligotrophic waters of the eastern Mediterranean, there are generally no reports of eutrophication or degraded water quality related to finfish or shellfish farms [45–47] and this was interpreted as the ability of the oligotrophic system to successfully assimilate the nutrients released by the farms. In an effort to understand whether nutrient release from aquaculture might have large-scale effects on the Mediterranean ecosystem, Karakassis et al. [48] employed a model to examine various production scenarios. They concluded that if aquaculture continues to grow and expand at present rates, farm wastes may increase overall nutrient (mainly N and P) levels by 1%, however, this is a general assessment and does not take into account localized effects. As suggested by Pitta et al. [49] in a study of three different Greek farms, it is likely that dispersion and dilution of the nutrients, combined with efficient herbivore grazing of algae (that develop from the released nutrients) were the reason for the absence of eutrophication around fish farms.

Although water quality is generally not affected, fish farms that operated over or near seagrass beds (especially *Posidonia oceanica*) exerted a clear effect on these [50, 51] and it was proposed that this may be related to the enhanced flux of dissolved and particulate nutrients from aquaculture. In an attempt to identify the effect of the plume of nutrients released from fish farms on water quality, Dalsgaard et al. [52] devised an innovative "bioassay" to measure the effect of dissolved nutrients released from fish farms on micro- and macro-algal production. They determined that primary productivity decreased with distance from the fish farms, yet by comparing bioassays with and without grazer exclusion, Pitta et al. [53] found that planktonic grazers (probably protista) play a key role in transferring nutrients up the food web.

Measures to Reduce Effects on the Water Column

One of the primary considerations when evaluating the suitability of sites for aquaculture is how they will interact with the surrounding marine system [54]. It does not pay, for example, to place net cage farms in shallow, poorly flushed waters (e.g., lagoons) because the organic and inorganic enrichment may affect both the marine ecosystem and the farmed organisms. Nevertheless, some farms have been deployed in unsuitable locations and these need to be relocated to allow the environment to recover and to enable the healthy growth of farmed finfish.

One of the early water quality problems associated with Mediterranean fish farms was the presence of an oily film around the cages. This was generally related to the large percentage of dust (pulverized feed pellets) in the pelleted food, which is not available to the farmed fish. Because this causes considerable loss to the farmers, and reduced water quality (stimulated bacterial growth also depletes the water of essential dissolved oxygen), the problem was rapidly addressed and most of the pelleted feeds are now extruded to improve pellet integrity and reduce feed loss and feed dust is collected and recycled.

A similar problem was identified in the tuna-penning industry. Unlike sea bream and bass that feed on formulated pellets, tuna are fed whole (preferably oily) fish such as sardine, anchovy, and mackerel. When these fish are offered to the tuna, the water around the pens often has an oily film and emits a strong smell. Moreover, in some cases, divers have complained of poor visibility near the pens. As described above, research is ongoing to develop artificial diets for tuna [55] that will address not only the problems related to feeding with fresh fish but also the water quality problems.

Disease

Intensive aquaculture systems are very susceptible to disasters such as loss of the farmed stock. Among the various causes of such disasters, disease outbreaks rank highest [56] and may lead to great losses within a very short period of time.

Most finfish cage farms in the Mediterranean are intensive, that is, they have high stocking density in order to be economically profitable and to compensate for the low profit margin of sea bream and sea bass, the main species reared in this region. Although cage stocking densities are usually <25 kg m^{-3}, in some farms stocking densities are higher and such conditions may cause a reduction in fish growth rates, suppression of immune mechanisms [57–59], and ultimately greater susceptibly to disease agents, including opportunistic bacterial and viral pathogens and eukaryotic parasites [60]. Current estimates of average mortalities for farmed sea bream and sea bass as a result of disease are 10% and 20%, respectively, for growth from juvenile to market size (350 g) fish. In many cases, the profit margin for these fish is not much higher than 10–20%, which has therefore obliged the aquaculture sector to consider various options to address this problem. Moreover, there is concern regarding the potential transmission of disease from the farmed stock to wild fish, based on studies of disease transfer among Atlantic salmon (e.g., [61]). It is noteworthy that although there are numerous examples of disease exchange between caged and wild fish (e.g., [62–64]) in the Mediterranean, and other seas, most of these are not clearly understood [65, 66] and their effect on native stocks is unclear.

Measures to Reduce Disease Outbreaks Numerous antibiotics have been tested against the common farmed fish diseases and there are currently treatments available for most bacterial fish pathogens [67]. However, the routine use of antibiotics in marine aquaculture is problematic and has declined for a number of reasons. First, as specified above, there are concerns related to human and environmental health and safety. Second, although many of these drugs work well in freshwater, some of the major antibiotics, such as quinolones and tetracyclines interact with the divalent cations that are abundant in seawater (mostly Mg^{2+} and Ca^{2+}) which massively reduces their function and efficacy [68, 69]. Moreover, there is no "harmonization" regarding antibiotics use among Mediterranean countries and the list of pharmaceuticals licensed for fish varies from country to country, complicating international trade and marketing.

In addition to bacterial pathogens, there are several parasitic diseases that may stunt growth rates, cause loss of fecundity, and even mortality in Mediterranean fish. These include various protozoa and metazoa, which are classified as ecto- and endoparasites according to their distribution on/in the fish. Pathologists consider the myxosporeans *Myxidium leei*, *Polysporoplasma sparis*, and *Ceratomyxa* sp., isopods, copepods, and monogenean infections among the more problematic parasites.

Athanassopoulou et al. [70] reviewed the drugs used against a variety of parasites and found that amprolium and sanilomycin were the most effective against myxosporans in cultivated breams. Moreover, extracts from oregano revealed anti-myxosporan as well as antibacterial properties. Ivermectin and deltamethrin – drugs used to combat sea lice, have also been tested against copepod and isopod infections in sea bass and were fairly effective, but they tend to become toxic to the fish at fairly low levels.

In order to limit the use of antibiotics and other chemotherapeutants, the European Union established the "Maximum Residue Limit" (MRL) regulation, which monitors the presence of these drugs in all agriculture and aquaculture products and this has had a dramatic effect on the use of therapeutants. Because the MRL differs among countries insofar as which compounds are regulated and which are not, there is a lot of work ahead, but despite this, the trend looks very promising.

Vaccines are one of the preferred measures for prevention of disease outbreaks, however because the Mediterranean finfish market is still fairly small, only a limited number of vaccines have been developed for commercial use. Moreover, consumer concerns and increasing restrictions regarding their use have led the industry to consider other alternatives to disease "management" [71, 72].

There are other alternatives to the use of chemotherapeutants and vaccines against disease. One of the key factors in the prevention of disease is good husbandry, which focuses on minimizing stress to the farmed stock. This includes proper stocking densities, optimal nutrition, sanitary practices, use of vaccines, and probiotics [66, 73]. The practice of good husbandry ensures fish are healthy and able to resist various disease agents naturally found in their environment. When they become stressed, the dietary requirements of fish for nutrients and vitamins change and a diet that compensates for such needs may optimize the growth of fish in captivity.

In recent years, it has become clear that the integrity of the gastrointestinal tract is essential in defense against pathogen attack as well as in proper endocrine and osmoregulatory activity. In recognition of this, Dimitroglou et al. [74] added the mannan oligosaccharide, Bio-Mos® (Alltech Inc, USA) to the diet of several marine fish including gilthead sea bream and found that this improved the gut morphology.

It is assumed that one of the roles that the mannan plays in protection of the fish is agglutination of pathogenic bacteria, which prevents their colonization of the gut. Indeed, the application of Bio-Mos significantly reduced the bacterial load in fish guts by reducing the biomass of aerobically cultivated bacteria [74]. Torrecillas et al. [75, 76] applied Bio-Mos to sea bass juvenile diets and found that it improved growth rates by 10%. Moreover, challenge trials using *Vibrio alginolyticus* showed that Bio-Mos fed sea bass had fewer of the pathogenic vibrio in their gut.

In recognition of the essential role of healthy gut flora in fish, especially in young fish, the use of immunostimulants, prebiotic, and/or probiotic bacteria have been proposed as a means to reduce gut colonization by pathogens [77], thereby improving the survival of cultured fish. Probiotics involves the addition of nonpathogenic bacteria to the diet and water of fish with the aim of loading the gut with bacteria that will prevent colonization by competing pathogens. The use of prebiotics and immunostimulants focuses on boosting the fish immune system so that the fish may more readily recognize and repel pathogen gut colonization. Although research has been conducted on the use of probiotics in Mediterranean aquaculture, (e.g., [78–80]), this approach has not successfully replaced the use of antibiotics to combat disease. One of the problems related to the use of probiotic bacteria is concern that these may not be as safe as they are supposed to be and their use may lead to other problems rather than a sustainable solution in the battle against disease.

Immunostimulants are commonly used in finfish farming to reduce the risk of disease by stimulating the protective activity of the immune system. The common forms of immunostimulants used in sea bream and sea bass aquaculture include ascorbic acid, a-tocopherol, and glucans [81, 82], which are added to the feed. Their presence appears to enhance antibacterial lysozyme activity and other indicators of disease resistance, but there is considerable discussion about their effectiveness due to the inherently wide range in concentrations and activities of the disease resistance molecules in fish serum.

Another approach to reduce the risk of disease is by means of classical selection/breeding for disease resistance by means of selective breeding programs [83]. The understanding of immune regulatory genes responsible for resistance to finfish pathogens is still in its infancy in Mediterranean aquaculture, but this field is rapidly expanding and it is anticipated that genetically superior lines will dominate the populations of fish reared in intensive aquaculture [84].

Escapes

In addition to problems related to disease and fluctuating profitability of aquaculture operations, fish farmers are also concerned with keeping their fish within the cages so that these can be marketed at the end of the growth cycle. There are many factors that may lead to loss of the farmed stock, including storms that may physically damage the net cages, predators (e.g., sharks, dolphins, bluefish, seals) that may bite the nets in their attempt to eat the enclosed fish, human error (e.g., during replacement of net cages or during harvest), poachers that cut the nets to catch fish, collision of ships with cage farms, etc. All of these generally result in the release of farmed fish to the surrounding environment, involving financial loss to the farmer and potential environmental problems related to genetic and ecological interactions of the escapees with the wild fish. At present, there are an estimated >1 billion fish; mostly sea bream and sea bass in net cage farms throughout the Mediterranean as compared to much smaller stocks of the wild populations of these species [85], so the potential impact of escapees is considerable. Because many Mediterranean countries do not require farmers to report escapes, there are no reliable data on the frequency of escapes, however, it is assumed that the percentages of escapees are similar to those reported in Norway [86], ranging between 0% and 6%. In addition to genetic "pollution" of the wild-stock gene pool, and potential competition between escapees and wild fish over the same habitat and food resources, there is also concern regarding the spread of disease from farmed fish to wild fish populations [87].

Measures to Reduce Escapes and Damage due to Escapes As the volume of aquaculture production increases in the Mediterranean to match demand, and with the anticipated addition of North Africa to the fish-producing countries, there is a growing need for regulation in order to minimize problems related to escapes. In order to appreciate the scale of escapes from Mediterranean aquaculture, there is a need to legislate reporting of escape events, as is currently done in other parts of the world. Moreover, several new finfish species have been domesticated and their potential effect, as escapees, on wild populations and on the ecosystem need to be assessed. In addition, in order to assess escape impacts, it is useful to be able to track the escaped fish, as described by Triantaphyllidis [88].

There are many measures that may be employed to reduce the risk of escapes from fish cages. Storm damage to farm systems is one of the major causes of escapes and employment of a reliable standard, as practiced in Norway (NS 9415 – requirements for design and operation of marine fish farms) is a promising approach to reduce such risks. Even sturdy, reliable cages are occasionally damaged by especially strong storms, but most of the surface wave energy is concentrated in the upper 10 or 15 m of the water column [85]. Submersible cage systems designed for open sea conditions, such as the Sub-flex system (www.subflex.org) and the Ocean-Spar system (www.oceanspar.com/) are an option to reduce mechanical stress to net cages in high-energy environments. Added advantages of submersible cage systems include the reduced risk of collisions with maritime vessels and the reduced visibility following the "out of sight–out of mind" solution to NIMBYism. Human poachers are a problem that may be reduced by vigilance and by cooperation with the local police or security forces. Marine predators that bite net cages from the outside may be deterred by using stronger materials, though this has financial consequences, or by embedding chemical deterrents in the net material. Several farmed species tend to bite the net material from the inside and this may create holes enabling escapes. The biting may be prevented by using taste deterrents, as described for predators, or stronger material that will be more bite resistant. Moe et al. [86] suggest making the cage environment more "appealing" or stimulating to reduce gnawing on the net mesh which they attribute to boredom.

In addition to reducing the risk and frequency of escapes, there is also a need to reduce the impacts caused by the escaped fish. One direction that is being tested is the development of sterile triploid sea bream and sea bass that will not be able to pass on their genes to wild fish. Another possibility is the recapture of the escaped fish, but this direction is still in early developmental stages. The location of fish farms relative to areas of high ecological sensitivity or to spawning grounds should be one of the major considerations in light of the possibility that some of the stocked fish may escape.

Introduced Exotic Species

Invasive species are probably the cause of the greatest ecological problems identified over the past century, not only in terrestrial but also in aquatic and marine systems [4]. This problem has intensified over the past 20–30 years, as the volume of intercontinental traffic has increased. Aquatic invasive species are a major threat to marine biodiversity and impact human health and the economy [89]. There are numerous examples of the impacts of invasives on human welfare and environmental health, for example, the invasion of the Black Sea by the exotic ctenophore *Mnemiopsis leydi*, which caused the collapse of most of the local fisheries [90]; invasion of the eastern Mediterranean by the Red Sea medusa, *Rhopilema nomadica*, which has heavily impacted Israeli and Turkish fisheries, tourism and coastal facilities [91].

In the eastern Mediterranean, exotic introductions are mainly channeled through the Suez Canal whereas most of the successful invaders in the western Mediterranean have been introduced by ships and via aquaculture [92].

Species introductions via aquaculture activities may be intentional or accidental, though the consequences are generally similar. Intentional introductions generally include the import of an exotic species and its release into the environment, without the intention that it spreads and dominates its new habitat. Examples include shellfish such as the Japanese oyster that was brought to France and spread rapidly throughout French coastal waters and certain species of sport fish that were intentionally released in northwestern US waters. The majority of introductions are not intentional but rather accidental and may occur in a number of ways. One common example of an accidental introduction is the transfer of a local species of oyster from a hatchery to the coast in a restocking program and the accidental release of an associated seaweed with the oysters. In another case, recreational boaters did not thoroughly wash the bottom of their boat after a holiday in a given bay and when they transported the boat back to their own shore, they brought with them a cryptic gastropod which subsequently invaded the new environment and decimated the local clam population.

Measures to Reduce the Invasion of Exotic Aquatic Species and Associated Damages In order to avoid the various risks involved in the use of exotic species, it is essential to rear/grow native species, as a rule. In many cases, the commercially attractive species are not native and farmers prefer to culture nonnative species. Introduced species may only be considered after taking all required precautions as specified in the ICES *Code of Practices on the Introductions and Transfers of Marine Organisms* [93] and the report on *Alien Species in Aquaculture* by Hewitt et al. [94]. Because the introduced species may escape and invade either local or neighboring environments, with implications for marine biodiversity, there is a need for both regional and international collaboration to address transboundary introductions and invasion issues, as discussed in UNEP [92].

The Mediterranean Aquaculture Market

The dominant species currently reared in the Mediterranean Sea are sea bream and sea bass [95]. These are native species that have been traditionally fished and eaten for centuries in many of the Mediterranean countries. Aquaculture has greatly increased the availability of these fish to the public and as production has increased, the price of the farmed fish has dropped dramatically so that in many cases its profitability is questionable. One of the important elements of a sector's sustainability is its economic performance yet the current trend in the Mediterranean is a plateau in profitability, that is, stagnation due to a glut in production of the two main species and a concurrent drop in their market value.

Alternative Aquaculture Species In order to survive and grow, the Mediterranean aquaculture sector needs to diversify its marine finfish production and include species with high market value. There are many native Mediterranean species that have a market because they are caught and sold by fishers and are suitable for cage culture. These include several species that have already been successfully reared in the eastern Mediterranean, such as Grey mullet (*Mugil cephalus*), Dover sole (*Solea solea*), Meagre (*Argyrosomus regius*), Sharp snout sea bream (*Diplodus puntazzo*), White bream (*Diplodus *sargus*), Red porgy (*Pagrus pagrus*), Shi drum (*Umbrina Cirossa*), Striped sea bream (*Lithognathus mormyrus*), Pandora (*Pagellus erythrinus*). Although these fish are commercially available for aquaculture, there are several bottlenecks that prevent large-scale production. These include lack of knowledge regarding their nutritional requirements, lack of farm facilities for production, slow growth rates (may be related to nutrition or other problems), sensitivity to certain pathogens.

Ecosystem Effects

It has been shown that Mediterranean fish farms generally have a local effect, primarily on the underlying benthos, as described above, yet within a short distance from the cages, this effect rapidly dissipates. It has been suggested that the large load of nutrients that pass via the farmed fish into the marine environment are rapidly processed by the biota, yet may exert some ecosystem effects. This hypothesis was tested by comparing the biological/chemical composition of seawater from fish-farming zones (within 2–3 nautical miles of fish farms) versus nonfarm zones (20 nautical miles of fish farms) in three parts of the Aegean sea in May and in September [49]. The data indicate that there is rapid transfer of nutrients up the food web, from the primary producers, via herbivores [53] to fish [96, 97]. These findings may be interpreted in a number of ways and their ramifications are debatable. If the precautionary approach is adopted, it is not clear what sort of implications these ecosystem-level changes may have and so they should be regarded with caution. On the other hand, if fish farms increase the size of natural fisheries, providing fishermen with an increased catch, this may be regarded as a positive externality of aquaculture (positive socioeconomic impact), which should be encouraged.

Seagrasses

One of the unique features in the Mediterranean Sea is the seagrass meadows of *Posidonia oceanica*. This slow-growing seagrass species occurs exclusively in the Mediterranean and grows best in clear, oligotrophic waters [98]. *P. oceanica* provides many ecosystem services, such as seabed stabilization, provision of a complex habitat to many larval and juvenile animals, oxygen production/release and long-term storage of CO_2 as plant tissue. Due to their slow growth rates, there is concern that these seagrass beds will not manage to recover if damaged and this important ecosystem and the services it provides may be lost. Marine botanists have calculated that some clonal colonies of *P. oceanica* may be 100,000 years old, that is, these are probably the largest and oldest-known living "organisms" on earth (http://en.wikipedia.org/wiki/Posidonia_oceanica). Because of their unique features, important ecological role and relatively low resilience to damage there is a strong movement in many Mediterranean countries to conserve and protect seagrass meadows from pollution, coastal development, trawling, and aquaculture. Recent work indicates that *P. oceanica* meadows located near or under fish farms have sustained considerable loss, including reduced meadow density, high shoot mortality rates (50-Diaz-Almela et al. 2008), increased epiphyte cover [99, 100] and very slow recovery rates following farm removal [101]. An analysis of several variables that may cause the observed damage to *P. oceanica*, in the context of the *MedVeg* project, has identified the deposition of particulate organic matter from the farms onto the seagrasses as the main factor leading to seagrass decline [102].

Measures to Protect Seagrass Meadows A set of recommendations were published by Pergent-Martini et al. [103] for the protection of Posidonia from fish farms, guided by the precautionary principle. These specified that: (a) Fish farms should not be situated directly over *P. oceanica* and *Cymodocea nodosa* (another important seagrass) meadows. (b) If seagrasses grow where a farm is planned, cages should be located at least 200 m from the nearest meadow. (c) Because these seagrasses generally occur at depths shallower than 45 m, farms should be set up at depths of 45–50 m where possible. (d) Environmental Impact Studies that relate to all seagrasses in the region should precede all lease requests to set up a fish farm. (e) If there are *P. oceanica* meadows near fish farms, these should be examined every 4 years to assure they have not been affected by the farming activity. On the basis of more recent findings, Holmer et al. [102] recommended to increase the distance between seagrass beds and fish farms to 400 m and to establish permanent seagrass plots to enable annual monitoring and sampling for seagrass health.

Future Directions

In the early 1990s, finfish aquaculture was generally a novelty in most parts of the Mediterranean, but this has changed radically during the past 20 years, as cage culture has spread throughout the region. Aquaculture is one of the fastest growing sectors worldwide and in the Mediterranean and it has many advantages over other food production industries, but in order to maintain a "green" image, aquaculture production and development must be sustainable. Progress has been made in many aspects of aquaculture technology but there are several areas that require attention and improvements in order to make this industry more environmentally and socioeconomically sustainable. Although numerous projects have focused on understanding the environmental interactions of aquaculture, the calculation of a reliable "carrying capacity" for aquaculture in a given water body is still generally beyond our means, that is, there is a need for further study of ecological processes on a variety of different scales with respect to fish farms. Because there are so many different types of habitats and ecosystems within the Mediterranean Sea (e.g., hard vs. soft seafloors, Adriatic vs. Levant, etc.), it is essential that the ecological and socioeconomic research address region-specific issues [45].

As aquaculture expands into new areas and new species, there is added urgency to improve the understanding of fish pathology in Mediterranean systems. In addition to bacterial diseases, there is a need for research into antihelminthic treatments, and better understanding of life cycles and early diagnostics for many of the Mediterranean parasites. In view of EU policies concerning reduction of chemical use in the aquatic environment, the prudent and effective use of chemotherapeutants is essential. This may be achieved by combining therapeutic treatments with such health management strategies as breeding of tolerant fish, improving water quality, and vaccination.

Escaped fish may impact wild fish through competition, predation, habitat displacement, gene pool dilution, etc. In an attempt to reduce the numbers of escapees, progress is being made (e.g., in the EU project "Prevent Escape," which includes several partners from Mediterranean countries) in the design of cages that should be more damage resistant and in devising strategies to track the escapees and to reduce migration away from the breeched cages.

A Need for Legislation

One of the areas that urgently requires attention to enable development of the sector is legislation since this aspect is inadequately addressed in many Mediterranean countries. Moreover, in many countries that are active in aquaculture, there is a policy vacuum with regard to this sector. There is a need for clear rules and standards for licensing, planning, environmental impact assessment (EIA), administrative organization, and coordination. In the absence of clarity and transparency in such matters, investors and entrepreneurs will not take the risks involved in establishing aquaculture operations and the development of the industry will be retarded and sluggish. In a review of the legal obstacles to aquaculture, Van Houtte [104] included: (a) the legal status of water used (public or privately owned), the nature of water used (marine, brackish, or freshwater); (b) the legal status and nature of the land used (coastal vs. inland; private vs. public); and (c) the need for government regulation of aquaculture, and related activities. Moreover, the lack of coordination among public and regulatory agencies with regard to the EIA process, planning, etc. complicates the aquaculture application process. To further complicate matters, the permit application process is complex, cumbersome and very time consuming. The number of laws, regulations, rules, and procedures involved in the application process is large and many different authorities are involved at several levels. On top of that, the application requirements vary widely from country to country and in some countries, aquaculture legislation may vary internally on a provincial or regional basis.

One of the most problematic policy issues has to do with site selection and site allocation for aquaculture. As an economic activity that takes place, and has an effect on the littoral, aquaculture competes with many other uses of the coastal zone and needs to be included in Mediterranean coastal planning and management schemes. In recognition of the rapidly growing sector, in 2002 the European Union acknowledged that planning and coastal management would be among the major challenges facing European aquaculture. This was reinforced by the recent EU [105] communication, which emphasizes that “area choice is crucial and spatial planning has a key role to play in providing guidance and reliable data for the location of an economic activity, giving certainty to investors, avoiding conflicts and finding synergies between activities and environments with the ultimate aim of sustainable development” and invites all Member States to “develop marine spatial planning systems, in which they fully recognize the strategic importance of aquaculture.”

One of the options chosen by some Mediterranean countries is zoning, that is, allocating a specific area for aquaculture as a means to reduce conflicts between coastal activities. In principle, this sort of approach simplifies things, provided: (a) the criteria used for selection of the aquaculture zones were appropriate and (b) the decisions regarding zoning involved the stakeholders and their interests. It is noteworthy that although there is aquaculture zoning in some countries, aquaculture jurisdiction generally falls under regional governance, that is, there are no national zoning plans in the Mediterranean [54]. Although zoning is probably one of the better options for site selection, the lack of national coordination regarding the allocation of space for aquaculture will probably increase conflicts with time, thereby jeapordizing the sustainability of the industry. It would therefore be prudent to promote national zoning policy for aquaculture in the Mediterranean.

The conflict over space is fierce in the coastal zone as there are many competing stakeholders and one of the solutions to this is to go offshore [95, 106]. There have been many initiatives over the past few decades promoting offshore or open-ocean aquaculture, including several international conferences in the Mediterranean; however, a number of obstacles have prevented the realization of this concept. These obstacles include (a) economic feasibility of such ventures; (b) engineering and technological solutions for aquaculture in sites exposed to oceanic conditions; (c) international and national (government) support for an offshore aquaculture industry; (d) investors willing to take the risks involved in offshore aquaculture; (e) lack of understanding of the ecological ramifications (water column and benthos; local and regional effects) of large-scale aquaculture in exposed sites; and (f) the biological effects of cultivation in exposed conditions (storms, currents, predators, etc.) on the farmed stock, and other similar issues. At present, there are a few Mediterranean fish farms situated in exposed, offshore sites, but these are the exception rather than the rule, and most farms are situated in protected or semi-sheltered sites. A move away from the coastal zone into offshore waters will probably become a reality rather than an option in the near future and the aquaculture sector stands to benefit if it can accept this and help establish the scientific basis and technology in advance.

Integrated Aquaculture

Another option that makes considerable ecological and economic sense is an integration of different forms of aquaculture within the same farm. By arranging systems for rearing finfish (a form of “fed” aquaculture) adjacent to systems for growing shellfish and/or seaweeds (extractive aquaculture), it may be possible to increase farm sustainability on a number of levels. On the ecological level, shellfish and algae are called “extractive” because they extract their nutrients or food from within the system (autochthonous), and can therefore help reduce the nutrient loads from fish farms. Finfish are usually “fed” with feed that is manufactured from materials that come from outside the system (allochthonous) and the release of wastes and uneaten feed from the farms may affect water and sediment quality and even cause eutrophication. On the social level, cultivation of different products as compared to monoculture will require greater manpower and expertise and create the opportunity for greater employment, both within the farms and in the form of support services. On the economic level, additional crops should increase farm profitability, provided the filtering organisms are able to absorb the nutrients efficiently and they fetch a good price at market. Moreover, by diversifying the cultured stock, the farmer protects himself from risks related to market fluctuations, storms, and disease. Integrated aquaculture is currently practiced in Canada and in China on pilot to commercial scales but it is not clear how this approach will develop with time. In the Mediterranean Sea, there are no commercial integrated aquaculture farms [21] and this is due to the fact that either the secondary crop is a low-value (not profitable) product or the secondary (extractive) crop is not able to grow in the oligotrophic conditions that characterize Mediterranean waters. The potential for integrated Mediterranean aquaculture exists, but it must be both ecologically and economically viable to work.

Herbivorous Fish

One of the major challenges for both global and Mediterranean aquaculture is the limited supply of essential fish oil and fish meal [107]. The artificial diets of many farmed fish, including salmon, sea bream, and sea bass rely heavily on fish meal and fish oil, which places considerable pressure on wild fisheries (the source of fish meal and oil), severely jeapordizing the sustainability of the sector [108]. Several strategies have been proposed to address this problem, including the extraction of oils from fish-processing wastes [109, 110] and from fishery by-catch discards (the noncommercial fish and animals that are caught by fishermen and subsequently thrown back to sea), and feeding fish with plant oils. There has been some success in the replacement of fish oils with plant oils [107], but many fish species have reduced survival and growth rates when reared without fish oils.

Another solution that has been proposed to address this problem is the rearing of herbivorous fish that do not require fish oils. Although these are generally not the highest value fish, they are nonetheless commercial species that are profitable to rear. The most common farmed herbivore in the Mediterranean is the diadromous gray mullet, *Mugil cephalus* (www.fao.org/fishery/culturedspecies/Mugil_cephalus/en). A lot of the pond rearing technology of this species was developed in Israel [111] and included polyculture. Egypt, the world leader in mullet production, has recently exceeded 1 million t/y. Although this fish is common in some of the southern Mediterranean countries, it does not have a large market in southern Europe and this is a challenge that needs to be overcome to promote herbivores as more sustainable species for aquaculture. Another problem that exists for *M. cephalus* is the absence of commercial hatcheries. Despite recent breakthroughs in spawning induction [112, 113], juvenile mullets are still collected from river mouths for aquaculture purposes thereby jeapordizing natural populations. These problems need to be addressed if this species is to be seriously considered a sustainable alternative to the common Mediterranean carnivores.

Indicators for Sustainable Aquaculture

The Water Framework Directive establishes the Environmental Quality Standards for European waters, and all activities that may affect environmental quality, for example, aquaculture must comply with these standards. Aquaculture lease applications generally include Environmental Impact Assessments (EIA), which assess risks and predict the impacts of aquaculture. Monitoring is an approach to test if EIA predictions were correct, and to establish a feedback system to protect both the environment and the fish farmer. The Modeling-Ongrowing fish farms-Monitoring (MOM) system [114, 115] was developed for salmon farming in Scandinavia, and includes a feedback process of EIA – monitoring – farm adjustment. Although the MOM concept was developed for Scandinavian farms, this approach has been adopted by the operators of several farms in the Mediterranean Sea to monitor their performance and environmental status. Monitoring generally includes measurement of: (a) physical variables, such as hydrography, weather, water temperature, sediment type, etc.; (b) chemical variables, including dissolved oxygen, nutrients, suspended solids, dissolved and particulate organic matter, etc.; and (c) biological attributes, for example, algal pigments, biomass, productivity, macrofauna abundance, diversity, etc. Fernandes et al. [116] reviewed the science underlying aquaculture monitoring in Europe and found that it was generally motivated by research interests rather than by clear environmental objectives. Whereas comprehensive monitoring of marine environments improves the understanding of the functioning of these systems [117], and thus the ability to predict the response of these waters to anthropogenic perturbations, it is often not necessary to include many of the variables that are monitored [102].

The CONSENSUS project recently established a set of 18 indicators (www.euraquaculture.info/index.php?option%20=%20com_content&task%20=%20view&id%20=%20149&Itemid%20=%20118) to promote "European Best Aquaculture Practice." These indicators are currently being evaluated to examine their practicality and suitability for the sector. In a separate project entitled *ECASA* (www.ecasa.org.uk/), a set of indicators to assess aquaculture–environment interactions were evaluated in order to streamline the farm monitoring process. This was done for aquaculture in both northern European and several Mediterranean countries (e.g., [118]) yet despite the advances made in that project, there is still a need to further streamline the list of indicators. The main criteria that should be used as a guideline in the quest for optimal indicators have been described in UNESCO [119] and include: (a) relevance, (b) feasibility (amount of effort, expertise, and cost required to obtain the data), (c) sensitivity (to inform on how the environment is responding), and (d) clarity (how easy it is for stakeholders to understand). Although progress has been made toward developing the final list of such indicators for aquaculture, this work is only partially done and further work is needed to achieve this.

Bibliography

1. FAO (2009) Integrated mariculture: a global review. In: Soto D (ed) FAO fisheries and aquaculture technical paper no 529. FAO, Rome, pp 133–183
2. Zveryaev II, Arkhipkin AV (2008) Structure of climatic variability of the Mediterranean sea surface temperature. Part I. Standard deviations and linear trends. Russ Meteorol Hydrol 33:377–382
3. Thingstad TF, Krom MD, Mantoura RFC, Flaten GAF, Groom S, Herut B, Kress N, Law CS, Pasternak A, Pitta P, Psarra S, Rassoulzadegan F, Tanaka T, Tselepides A, Wassmann P, Woodward EMS, Riser CW, Zodiatis G, Zohary T (2005) Nature of phosphorus limitation in the ultraoligotrophic eastern Mediterranean. Science 309:1068–1071
4. EAA (European Environmental Agency), (2006) Priority issues in the Mediterranean environment. EEA Report 4:88
5. Abdulla A, Gomei M, Maison E, Piante C (2008) Status of marine protected areas in the Mediterranean Sea. IUCN, Malaga, p 152
6. Basurco B (2000) Offshore mariculture in Mediterranean countries. In: Muir J, Basurco B (eds) Mediterranean offshore mariculture. Options Méditerranées: Série B. Etudes et Recherches 30, October 20–24 1997, Zaragoza, Spain
7. Bardach JE, Ryther JH, McLarney WO (1972) Aquaculture: the farming and husbandry of freshwater and marine organisms. Wiley-Interscience, New York, p 868
8. Basurco B, Lovatelli A (2003) The aquaculture situation in the Mediterranean sea. In: International conference on sustainable development of the Mediterranean and Black sea environment, Predictions for the Future, Thessalonica, Greece, May 29–31 2003. http://www.oceandocs.org/bitstream/1834/543/1/Basurco.pdf
9. Nun M (1964) Ancient Jewish fishery (in Hebrew). Hakibbutz Hameuchad, Merkhavia, Israel
10. Raban A, Galili E (1985) Recent maritime archaeological research in Israel-a preliminary report. Int J Naut Archeol 14:321–356
11. Cataudella S (1996) Description of main Mediterranean aquaculture systems. Notes from the TECAM advanced course on food and feeding of farmed fish and shrimp, CIHEAM, FAO, NIOF, Alexandria, Egypt
12. Lovatelli A (2005) Report of the third meeting of the ad hoc GFCM/ICCAT working group on sustainable bluefin tuna farming/fattening practices in the Mediterranean. FAO fisheries report no. 779, FAO, Rome, p 108
13. FAO (1997) FAO technical guidelines for responsible fisheries. Fisheries department, aquaculture development. Report no. 5, FAO, Rome, p 40
14. Fish Site (2006) Farmed vs wild salmon? – a comparison. (http://www.thefishsite.com/articles/107/farmed-vs-wild-salmon-a-comparison)
15. Altintzoglou T, Verbeke W, Vanhonacker F, Luten J (2010) The image of fish from aquaculture among Europeans: impact of exposure to balanced information. J Aquat Food Prod Technol 19(2):103–119
16. Mazur NA, Curtis AL (2006) Risk perceptions, aquaculture, and issues of trust: lessons from Australia. Soc Nat Resour 19:791–808
17. Mazur N, Curtis A (2008) Understanding community perceptions of aquaculture: lessons from Australia. Aquac Int 16:601–621
18. Katrandis S, Nitsi E, Vakrou A (2003) Social acceptability of aquaculture development in coastal areas; the case of two Greek islands. Coast Manage 31:37–53
19. Korchenkov I (2010) Mariculture development: public policy, citizen's attitudes and factors affecting them. MA thesis, University of Haifa, Haifa, p 63
20. Whitmarsh D, Wattage P (2006) Public attitudes towards the environmental impact of salmon aquaculture in Scotland. Europ Environ 16:108–121
21. Angel DL, Freeman S (2009) Integrated aquaculture (INTAQ) as a tool for an ecosystem approach to the marine farming sector in the Mediterranean Sea. In: Soto D (ed) Integrated marine aquaculture: a global review. FAO fisheries and aquaculture technical paper no. 529, FAO, Rome, p 133–183
22. Brown JR, Gowen RJ, McLusky DS (1987) The effect of salmon farming on the benthos of a Scottish sea loch. J Exp Mar Biol Ecol 109:39–51
23. Gowen RJ, Bradbury NB (1987) The ecological impact of salmonid farming in coastal waters: a review. Annu Rev Oceanogr Mar Biol 25:563–575
24. Weston DP (1990) Qualitative examination of macrobenthic community changes along an organic enrichment gradient. Mar Ecol Prog Ser 61:233–244
25. Silvert WL (1992) Assessing environmental impacts of finfish aquaculture in marine waters. Aquaculture 107:67–71
26. Hargrave BT (ed) (1994) Modelling benthic impacts of organic enrichment from marine aquaculture. Canadian technical report, fisheries and aquatic science. Report # 1949, DFO, Canada, p 125
27. Holmer M, Kristensen E (1996) Seasonality of sulfate reduction and pore water solutes in a marine fish farm sediment: the

importance of temperature and sedimentary organic matter. Biogeochemistry 32:15–39

28. Pearson TH, Black KD (2001) The environmental impacts of marine fish cage culture. In: Black KD (ed) Environmental impacts of aquaculture. Sheffield Academic, Sheffield, pp 1–31
29. Karakassis I, Eleftheriou A (1997) The continental shelf of Crete: structure of macrobenthic communities. Mar Ecol Prog Ser 160:185–196
30. Karakassis I, Eleftheriou A (1998) The continental shelf of Crete: the benthic environment. PSZNI: Mar Ecol 19:263–277
31. Duineveld GCA, Tselepides A, Witbaard R, Bak RPM, Berghuis EM, Nieuwland G, van der Weele J, Kok A (2000) Benthic-pelagic coupling in the oligotrophic Cretan sea. Prog Oceanogr 46:457–481
32. Karakassis I, Tsapakis M, Hatziyanni E, Papadopoulou K-N, Plaiti W (2000) Impact of cage farming of fish on the seabed in three Mediterranean coastal areas. ICES J Mar Sci 57: 1462–1471
33. Vizzini S, Mazzola A (2004) Stable isotope evidence for the environmental impact of a land-based fish farm in the western Mediterranean. Mar Pollut Bull 49:61–70
34. Vizzini S, Savona B, Caruso M, Savona A, Mazzola A (2005) Analysis of stable carbon and nitrogen isotopes as a tool for assessing the environmental impact of aquaculture: a case study from the western Mediterranean. Aquac Int 13:157–165
35. Sarà G, Lo Martire M, Buffa G, Mannino AM, Badalamenti F (2007) The fouling community as an indicator of fish farming impact in Mediterranean. Aquac Res 38:66–75
36. Cromey CJ, Black KD (2005) Modelling the impacts of finfish aquaculture. In: Hargrave BT (ed) Environmental effects of marine finfish aquaculture. Springer, Berlin, Heidelberg, pp 129–156
37. Askens A, Izquiredo MS, Robaina L, Vegara JM, Montero D (1997) Influence of fish meal quality and feed pellet on growth, feed efficiency and muscle composition in gilthead sea bream (*Sparus aurata*). Aquaculture 153:251–261
38. Lupatsch I, Kissil GWm, Sklan D (2001) Optimization of feeding regimes for European seabass *Dicentrarchus labrax*: a factorial approach. Aquaculture 202:289–302
39. Clarke S, Smart A, van Barneveld R, Carter C (1997) The development and optimization of manufactured feeds for farmed southern bluefin tuna. Austasia Aquacult 11:59–62
40. McKindsey CW, Thetmeyer H, Landry T, Silvert W (2006) Review of recent carrying capacity models for bivalve culture and recommendations for research and management. Aquaculture 261:451–462
41. Porello S, Lenzi M, Persia E, Tomassetti P, Finoia MG (2003) Reduction of aquaculture wastewater eutrophication by phytotreatment ponds system: I Dissolved and particulate nitrogen and phosphorus. Aquaculture 219:515–529
42. Lenzi M, Gennaro P, Mastroianni A, Mercatali I, Persia E, Roffilli R, Solari D, Tomassetti P, Porrello S (2009) Improvement of a system for treating land-based fish-farm effluents. Chem Ecol 25:247–256
43. Sorokin YI, Sorokin PU, Ravagnan G (2006) Hypereutrophication events in the Ca'Pisani lagoons associated with intensive aquaculture. Hydrobiologia 571:1–15
44. Tett P, Gilpin L, Svendsen H, Erlandsson CP, Larsson U, Kratzer S, Fouilland E, Janzen C, Lee J-Y, Grenz C, Newton A, Ferreira JG, Fernandes T, Scory S (2003) Eutrophication and some European waters of restricted exchange. Cont Shelf Res 23:1635–1671
45. Karakassis I (2001) Aquaculture and coastal marine biodiversity. Oceanis 24:272–286
46. Danovaro R, Gambi C, Luna GM, Mirto S (2004) Sustainable impact of mussel farming in the Adriatic Sea (Mediterranean Sea): evidence from biochemical, microbial and meiofaunal indicators. Mar Pollut Bull 49:325–333
47. Sara G (2007) Ecological effects of aquaculture on living and non-living suspended fractions of the water column: a meta-analysis. Water Res 41:3187–3200
48. Karakassis I, Pitta P, Krom MD (2005) Contribution of fish farming to the nutrient loading of the Mediterranean Sea. Sci Mar 69:313–321
49. Pitta P, Apostolaki ET, Tsagaraki T, Tsapakis M, Karakassis I (2006) Fish farming effects on chemical and microbial variables of the water column: a spatio-temporal study along the Mediterranean Sea. Hydrobiologia 563:99–108
50. Diaz-Almela E, Alvarez E, Santiago R, Marba N, Holmer M, Grau T, Danovaro R, Argyrou M, Karakassis Y, Duarte CM (2008) Benthic input rates predict seagrass (*Posidonia oceanica*) fish farm-induced decline. Mar Pollut Bull 56:1332–1342
51. Apostolaki ET, Marba N, Holmer M, Karakassis I (2009) Fish farming enhances biomass and nutrient loss in *Posidonia oceanica* (L.) Delile. Estuar Coast Shelf Sci 81:390–400
52. Dalsgaard T, Krause-Jensen D (2006) Monitoring nutrient release from fish farms with macroalgal and phytoplankton bioassays. Aquaculture 256:302–310
53. Pitta P, Tsapakis M, Apostolaki ET, Tsagaraki T, Holmer M, Karakassis I (2009) "Ghost nutrients" from fish farms are transferred up the food web by phytoplankton grazers. Mar Ecol Prog Ser 374:1–6
54. IUCN (2009) Aquaculture site selection and site management . Guide for the sustainable development of Mediterranean aquaculture. IUCN Gland, Switzerland and Malaga, Spain, p 303
55. Mourente G, Tocher DR (2009) Tuna nutrition and feeds: current status and future perspectives. Rev Fish Sci 17: 373–390
56. Meyer FP (1991) Aquaculture disease and health management. J Anim Sci 69:4201–4208
57. Bonga SEW (1997) The stress response in fish. Physiol Rev 77:591–625
58. Wedemeyer GA (1997) Effects of rearing conditions on the health and physiological quality of fish in intensive culture. In: Iwama GK, Pickering AD, Sumpter JD, Schreck CB (eds) Fish stress and health in aquaculture, vol 62, Society for experimental biology seminar, series. Cambridge University Press, Cambridge

59. Pickering AD (1998) Stress responses of farmed fish. In: Black KD, Pickering AD (eds) Biology of farmed fish. CRC, Boca Raton, pp 222–255
60. Varvarigos P (1997) Marine fish diseases in Greece. Fish Farmer 20:10–12
61. Waknitz FW, Tynan TJ, Nash CE, Iwamoto RN, Rutter LG (2002) Review of potential impacts of Atlantic salmon culture on puget sound chinook salmon and hood canal summer-run chum salmon evolutionarily significant units. US Department of Commerce, NOAA technical memoranda, NMFS-NWFSC-53, p 83
62. Diamant A, Colorni A, Ucko M (2007) Parasite and disease transfer between cultured and wild coastal marine fish. In: Briand F (ed) Impact of mariculture on coastal ecosystems, CIESM workshop monographs, Lisbon no. 32, pp 49–54
63. Breuil G, Mouchel O, Fauvel C, Pepin JF (2001) Sea bass *Dicentrarchus labrax* nervous necrosis virus isolates with distinct pathogenicity to sea bass larvae. Dis Aquat Organ 45:25–31
64. Diaz Lopez B (2006) Bottlenose dolphin *Tursiops truncatus* predation on a marine fin fish farm: some underwater observations. Aquat Mamm 32:305–310
65. Diamant A, Colorni A, Ucko M (2007) Parasite and disease transfer between cultured and wild coastal marine fish. Impact of mariculture on coastal ecosystems. CIESM workshop monographs no: 32, Monaco, pp 49–54
66. IUCN (2007) Interactions between aquaculture and the environment. Guide for the sustainable development of Mediterranean aquaculture, Gland, Switzerland and Málaga, Spain, p 110
67. Austin B, Austin DA (1999) Bacterial fish pathogens: disease of farmed and wild fish. Springer-Praxis, Chichester
68. Barnes AC, Hastings TS, Aymes SGB (1995) Aquaculture antibacterials are antagonized by sea-water. J Fish Dis 18:463–465
69. Smith P, Niland T, O'Domhnaill F, O'Tuathaigh G, Hiney M (1996) Influence of marine sediments and divalent cations on the activity of oxytetracycline against *Listonella anguillarum*. Bull Eur Assn Fish P 16:54–57
70. Athanassopoulou F, Pappas IS, Bitchava K (2009) An overview of the treatments for parasitic disease in Mediterranean aquaculture. Options Méditerranéennes 86:65–82
71. Hansen GH, Olafsen JA (1999) Bacterial interactions in early life stages of marine cold water fish. Microb Ecol 38:1–26
72. Verschuere L, Rombaur G, Sorgeloos P, Verstraete W (2000) Probiotic bacteria as biological control agents in aquaculture. Microbiol Mol Biol Rev 64:655–671
73. Lall SP, Lewis-McCrea LM (2007) Role of nutrients in skeletal development in fish – an overview. Aquaculture 267:3–19
74. Dimitroglou A, Davies S, Moate R, Spring P, Sweetman J (2007) The beneficial effect of Bio-Mos on gut integrity and enhancement of fish health. Alltech technical seminar series, Dublin, November 2007
75. Torrecillas S, Makol A, Caballero MJ, Montero D, Robaina L, Real F, Sweetman J, Tort L, Izquierdo MS (2007) Immune stimulation and improved infection resistance in European sea bass (*Dicentrarchus labrax*) fed mannan oligosaccharides. Fish Shellfish Immunol 23:969–981
76. Torrecillas S, Caballero MJ, Sweetman J, Makol A, Izquierdo MS (2007b) Effects of feeding Bio- Mos on European sea bass (*Dicentrarchus labrax*) juvenile culture. Alltech technical seminar series, Dublin, November 2007
77. Gatesoupe J (2005) Probiotics and prebiotics for fish culture, at the parting of the ways. Aqua Feeds: Formulation & Beyond 2:1–5
78. Gatesoupe FJ (1999) The use of probiotics in aquaculture. Aquaculture 180:147–165
79. Picchietti S, Mazzini M, Taddei AR, Renna R, Fausto AM, Mulero V, Carnevali O, Cresci A, Abelli L (2007) Effects of administration of probiotic strains on GALT of larval gilthead seabream: immunohistochemical and ultrastructural studies. Fish Shellfish Immunol 22:57–67
80. Avella MA, Olivotto I, Silvi S, Place AR, Carnevali O (2010) Effect of dietary probiotics on clownfish: a molecular approach to define how lactic acid bacteria modulate development in a marine fish. Am J Physiol Regul Integr Comp Physiol 298:359–371
81. Jeney G, Galeotti M, Volpatti D (1994) Effect of immunostimulation on the non specific immune response of sea bass *Dicentrarchus labrax*. In: International symposium on aquatic animal health: program and abstracts, Seattle, Washington DC, September 4–8 1994, p 76
82. Bagni M, Archetti L, Amadori M, Marino G (2000) Effect of long-term oral administration of an immunostimulant diet on innate immunity in sea bass (*Dicentrarchus labrax*). J Vet Med 47:745–751
83. Knibb W, Gorshkova G, Gorshkov S (1997) Selection for growth in the gilthead seabream (*Sparus aurata*). Isr J Aquacult/Bamidgeh 49:57–66
84. Chistiakov DA, Hellemans B, Volckaert FAM (2007) Review on the immunology of European sea bass *Dicentrarchus labrax*. Vet Immunol Immunopathol 117:1–16
85. Dempster T, Moe H, Fredheim A, Jensen Q, Sanchez-Jerez P (2007) Escapes of marine fish from sea-cage aquaculture in the Mediterranean Sea: status and prevention. In: Briand F (ed) Impact of mariculture on coastal ecosystems, CIESM workshop monographs, no. 32, Lisboa, pp 55–60
86. Moe H, Dempster T, Sunde LM, Winther U, Fredheim A (2007) Technological solutions and operational measures to prevent escapes of Atlantic Cod (*Gadus morhua)* from sea-cages. Aquac Res 38:91–99
87. Diamant A (1997) Fish to fish transmission of a marine myxosporean. Dis Aquat Organ 30:99–105
88. Triantafyllidis A (2007) Aquaculture escapes: new DNA based monitoring analyses and application on seabass and seabream. In: Briand F (ed) Impact of mariculture on coastal ecosystems, CIESM Workshop Monographs, no. 32, Lisboa, pp 67–72
89. Carlton JT (2009) Deep invasion ecology and the assembly of communities in historical time. In: Rilov G, Crooks JA (eds) Biological invasions in marine ecosystems. Springer, Berlin, Heidelberg, pp 13–56

90. Shiganova TA (1998) Invasion of the Black Sea by the ctenophore *Mnemiopsis leidyi and* recent changes in pelagic community structure. Fish Oceanogr 7:305–310
91. Galil BS (2007) Seeing red: alien species along the Mediterranean coast of Israel. Aquat Invasions 2:281–312
92. United Nations Environment Programme (UNEP)/ Mediterranean Action Plan (2005) Action plan concerning species introductions and invasive species in the Mediterranean Sea. RAC/SPA, Tunis, p 30
93. ICES (2005) ICES code of practice on the introductions and transfers of marine organisms, p 30
94. Hewitt CL, Campbell ML, Gollasch S (2006) Alien species in aquaculture. Considerations for responsible use. IUCN, Gland, Switzerland and Cambridge, p 32
95. Cardia F, Lovatelli A (2007) A review of cage aquaculture: the Mediterranean Sea. In: Halwart M, Soto D, Arthur JR (eds) Cage aquaculture – regional reviews and global overview. FAO fisheries technical paper no. 498, FAO, Rome, pp 159–190
96. Machias A, Karakassis I, Labropoulou M, Somarakis S, Papadopoulou KN, Papaconstantinou C (2004) Changes in wild fish assemblages after the establishment of a fish farming zone in an oligotrophic marine ecosystem. Estuar Coast Shelf Sci 60:771–779
97. Machias A, Karakassis I, Giannoulaki M, Papadopoulou N, Smith CJ, Somarakis S (2005) Response of demersal fish communities to the presence of fish farms. Mar Ecol Prog Ser 288:241–250
98. Holmer M, Perez M, Duarte CM (2003) Benthic primary producers – a neglected environmental problem in Mediterranean maricultures? Mar Pollut Bull 46:1372–1376
99. Balata D, Nesti U, Piazzi L, Cinelli F (2007) Patterns of spatial variability of seagrass epiphytes in the north-west Mediterranean Sea. Mar Biol 151:2025–2035
100. Pérez M, García T, Ruíz JM, Invers O (2008) Physiological responses of the seagrass *Posidonia oceanica* as indicators of fish farm impact. Mar Pollut Bull 56:869–879
101. Delgado O, Ruiz JM, Pérez M, Romero J, Ballesteros E (1999) Effects of fish farming on seagrass (*Posidonia oceanica*) in a Mediterranean bay: seagrass decline after organic loading cessation. Oceanol Acta 22:109–117
102. Holmer M, Hansen P-K, Karakassis I, Borg JA, Schembri PJ (2008) Monitoring of environmental impacts of marine aquaculture. In: Holmer M, Black K, Duarte CM, Marbà N, Karakassis I (eds) Aquaculture in the ecosystem. Springer, Berlin, pp 47–86
103. Pergent-Martini C, Boudouresque CF, Pasqualini V, Pergent G (2006) Impact of fish farming facilities on *Posidonia oceanica* meadows: a review. Mar Ecol 27:310–319
104. Van Houtte A (2001) Establishing legal, institutional and regulatory framework for aquaculture development and management. In: Subasinghe RP, Bueno P, Phillips MJ, Hough C, McGladdery SE, Arthur JE (eds) Aquaculture in the third millennium. Technical proceedings of the conference on aquaculture in the third millennium, NACA and FAO, Bangkok, Thailand, February 20–25 2000
105. EU (European Union) (2009) Building a sustainable future for aquaculture: a new impetus for the strategy for the sustainable development of European aquaculture. Communication from the Commission to the European Parliament and the Council COM (2009), 162, Brussels, p 13
106. Stickney RR (1997) Offshore Mariculture. In: Bardach JE (ed) Sustainable aquaculture. John Wiley & Sons, New York, pp 53–86
107. Turchini GM, Torstensen BE, Ng W-K (2009) Fish oil replacement in finfish nutrition. Rev Aquacult 1:10–57
108. Naylor RL, Hardy RW, Bureau DP, Chiu A, Elliott M, Farrell AP, Forster I, Gatlin DM, Goldburg RJ, Hua K, Nichols PD (2009) Feeding aquaculture in an era of finite resources. Proc Natl Acad Sci 106:15103–15110
109. Raffi SM, (2006) Sustainable Utilisation of Bycatch Resources. In: Kannaiyan S, Balasubramanian T, Ajmalkhan S, Venkataraman K (eds) Biodiversity and Conservation of Marine Bioresources, National Biodiversity Authority, pp 107–113
110. Rubin, S (1993) Fish waste in Japan. Stiftelsen rubin, report no. 007/15, p 22
111. Sarig S (1981) The Mugilidae in polyculture in fresh and brackish water fishponds. In: Oren OH (ed) Aquaculture of grey mullets. Cambridge University Press, Cambridge, UK, pp 391–409
112. De Monbrison D, Tzchori I, Holland MC, Zohar Y, Yaron Z, Elizur A (1997) Acceleration of gonadal development and spawning induction in the Mediterranean grey mullet. Mugil cephalus: preliminary studies. Isr J Aquacult/Bamidgeh 49:214–221
113. Aizen J, Meiri I, Tzchori I, Levavi-Sivan B, Rosenfeld H (2005) Enhancing spawning in the grey mullet (*Mugil cephalus*) by removal of dopaminergic inhibition. Gen Comp Endocrinol 142:212–221
114. Ervik A, Hansen PK, Aure J, Stigebrandt A, Johannessen P, Jahnsen T (1997) Regulating the local environmental impact of intensive marine fish farming I. The concept of the MOM system (Modelling-Ongrowing fish farms-Monitoring). Aquaculture 158:85–94
115. Hansen PK, Ervik A, Schaanning M, Johannessen P, Aure J, Jahnsen T, Stigebrandt A (2001) Regulating the local environmental impact of intensive marine fish farming II. The monitoring programme in the MOM system (Modelling-Ongrowing fish farms-Monitoring). Aquaculture 194: 75–92
116. Fernandes TF, Eleftheriou A, Ackefors H, Eleftheriou M, Ervik A, Sanchez-Mata A, Scanlon T, White P, Cochrane S, Pearson TH, Read PA (2001) The scientific principles underlying the monitoring of the environmental impacts of aquaculture. J Appl Ichthyol 17:181–193
117. Tett P (2008) Fishfarm wastes in the ecosystem. In: Holmer M, Black K, Duarte CM, Marbà N, Karakassis I (eds) Aquaculture in the ecosystem. Springer, Berlin, pp 1–46
118. Borja Á, Germán Rodríguez J, Black K, Bodoy A, Emblow C, Fernandes TF, Forte J, Karakassis I, Muxika I, Nickell TD,

Papageorgiou N, Pranovi F, Sevastou K, Tomassetti P, Angel D (2009) Assessing the suitability of a range of benthic indices in the evaluation of environmental impact of fin and shellfish aquaculture located in sites across Europe. Aquaculture 293:231–240

119. UNESCO (2003) A reference guide on the use of indicators for integrated coastal management – integrated coastal area management dossier 1, Intergovernmental Oceanographic Commission Manuals and Guides No. 45.

Marine Biogeochemistry

Walker O. Smith, Jr.[1], Eileen E. Hofmann[2],
Anna Mosby[1]
[1]Department of Biological Sciences, Virginia Institute of Marine Science, College of William & Mary, Gloucester Pt, VA, USA
[2]Center for Coastal Physical Oceanography, Old Dominion University, Norfolk, VA, USA

Article Outline

Glossary

Autotrophic Organisms whose mode of nutrition is photosynthesis.

Biogeochemistry The biological and chemical processes that transform and cycle elements over various time and space scales and that determine the composition of the environment.

Biological pump The biological processes and transformations that move carbon from the surface to depth.

Cyanobacteria Prokaryotic phytoplankton.

Diatom Phytoplankton which are encased in frustule consisting of silica.

Euphotic zone The surface layer of the ocean where most primary production occurs, generally considered to be the depth to which 1% of surface radiation penetrates.

Heterotrophic Organisms who require reduced organic carbon as an energy and carbon source.

Nutrient Element that is required for biological activity and growth.

Oxidation Chemical reaction in which reactant loses electrons; half-reaction paired with reduction.

Photosynthesis The process by which radiant energy from the sun is transformed into chemical energy that can later be used to reduce carbon dioxide to organic sugars, which in turn are coupled to biochemical pathways to produce all compounds necessary for cell growth.

Phytoplankton Microscopic, often unicellular, floating autotrophs that live in the ocean's surface layer and form the base of nearly all marine food webs.

Reduction Chemical reaction in which reactant gains electrons; half-reaction paired with oxidation.

Definition of the Subject and Its Importance

The biogeochemistry of the world oceans has been studied for many decades, and major advances in understanding have been linked with development of new techniques and tools that allow the accurate representation of various organic and inorganic pools within the water. The classic study of Redfield [1] showed that some critical bioactive compounds (carbon, nitrogen, phosphorus, oxygen) occur in particular ratios to one another that are relatively invariant over space and time and provided a description of the relationship between the ratio of nitrogen to phosphorus (N:P) for inorganic and plankton pools. The processes that control these compounds were assessed, and it was concluded that phosphorus concentrations are largely controlled by terrestrial inputs, whereas nitrogen is under biological control.

Subsequent studies have provided more detailed investigations of the processes controlling these ratios. These studies benefited from the development and standardization of methods for accurately measuring dissolved organic carbon (DOC) and dissolved organic nitrogen (DON). The improved methodology, mostly developed during the 1980s, allowed the spatial (vertical and horizontal) and temporal changes of both DOC and DON to be quantitatively described.

Recognition of the importance of the flux of organic carbon to depth in mediating the marine response to increased atmospheric carbon dioxide concentrations stimulated development of technical approaches and instruments for assessing and quantifying the biological pump. This component of marine biogeochemical cycles is still a poorly constrained component of numerical models developed for simulation of ocean carbon cycling and climate, and technological approaches that result in better assessment of the flux of organic matter to depth continue to be developed and refined. Also, numerical models of biogeochemical processes are providing insights into critical processes and provide frameworks that allow measurements to be projected over larger space and timescales. Continued measurement and modeling of oceanic biogeochemical cycles is essential for understanding and projecting responses to natural and anthropogenic-induced climate change.

Introduction

The ocean is the dominant surface feature that has controlled much of the evolution, distribution, and success of life on earth, and the changes in ocean chemistry reflect the interaction with biota throughout geological time. The oceans were originally anoxic, but the evolution of organisms with oxygen-generating processes (photosynthesis) resulted in the conversion of the oceans to an oxygenated environment, which greatly altered the availability of some elements for those organisms. The cycling of elements within the earth's oceans and the complex relationships among the biological, chemical, and geological processes are the core of the study of marine biogeochemistry. Understanding these relationships is difficult and is further complicated by the space and time variability of the dominant processes that control the cycling of the different elements. Understanding the interactions and linkages among and between the cycles of biogeochemical elements is critically important for assessing and projecting the nature, degree, and direction of changes in ocean processes that may result from changes induced by natural and/or anthropogenic activities.

Elements in the ocean have characteristic vertical and horizontal distributions that result from the processes that regulate their long-term source/sink relationships. For example, oceanic carbon dioxide (CO_2) distributions are characterized by a horizontal concentration gradient that increases from the equator to the poles, which results from the greater dissolution of CO_2 in colder water. Carbon dioxide concentrations generally increase with depth due to remineralization in the deeper, older waters relative to its removal at the surface. Other elements may be controlled by different factors (e.g., sources from the sediments or hydrothermal vents; atmospheric sources) and have different vertical and horizontal patterns, but all interact to create the observed vertical distributions in the ocean. Understanding marine biogeochemistry requires knowledge not only of specific processes regulating a particular element, but also an understanding of the interdisciplinary aspects that control these cycles.

Nutrients are the biogeochemical elements that are required for biological activity. Some elements are greatly reduced in their concentrations by chemical or biological processes and can reach such low concentrations that they subsequently limit the growth of organisms in the sea. Such elements are thought of as limiting nutrients in the sense of the German agricultural chemist, Justus von Liebig, who suggested that the growth of plants is limited not by the total amount of resources, but by the resource in lowest abundance relative to the others.

Plant growth in the ocean is known to be limited by a small number of nutrients that include nitrogen, phosphorus, iron, silicic acid, and inorganic carbon. The cycling and processes that control the concentrations of these limiting nutrients are critical in the regulation of carbon cycling in the ocean, and hence their study forms the basis for most biogeochemical research.

The biogeochemical cycles described in subsequent sections use carbon as a "common denominator." Carbon is the basic component of organic matter, and with the advent of industrialization is being added to the atmosphere at an unprecedented, rapid rate, which is changing atmospheric temperatures and impacting the thermal equilibrium of the ocean. Also, carbon is absorbed from the atmosphere at the ocean surface where it reacts with ocean water to produce carbonic acid, thereby making ocean waters more acidic (reducing the pH), which has profound impacts on oceanic chemistry and biological activity. Thus, the production and oxidation of organic matter in the ocean has numerous critical interactions with all other elemental cycles, and is a major regulator of all marine biogeochemical cycles.

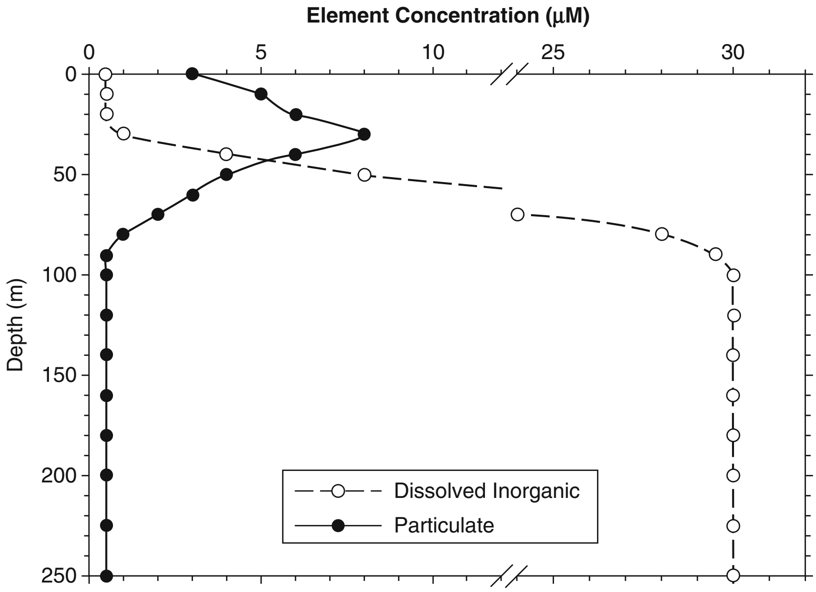
The Biogeochemical Cycle of Carbon

Carbon is the primary building block for all life because of its chemical ability to form a myriad of covalent bonds with itself and numerous other elements. As a result, the numerous complex organic compounds that form the basis of life systems are based on carbon. In the present-day ocean, synthesis of organic molecules (photosynthesis) is done largely by phytoplankton, which converts the radiant energy from the sun into chemical energy in the form of adenosine triphosphate (ATP). The ATP, along with reductant, is used to reduce CO_2 into simple sugars, which are in turn modified into all of the compounds required for cellular metabolism, growth, and division. Photosynthesis is dependent on energy from the sun, thereby confining this process to the euphotic zone, which is the part of the upper water column that receives at least 1% of the irradiance that reaches the sea surface. Phytoplankton require energy for the uptake and assimilation of nearly all elements. This dependence on light generally results in vertical distributions of nutrients that are characterized by reduced concentrations in the euphotic zone, where photosynthesis and growth are most active, and increased concentrations at depth, where photosynthesis and growth are reduced or absent (Fig. 1). This vertical profile is a typical of nutrient distributions throughout the oceans. The organic matter generated by photosynthesis and growth has roughly an inverse relationship to that of the inorganic building blocks (Fig. 1).

Redfield [1] suggested that organic matter (carbon, C) production in the sea occurs in relatively constant elemental ratios given by the relationship:

$$106CO_2 + 16H^+ + 16NO_3^- + H_3PO_4 + 122H_2O \leftrightarrow (CH_2O)_{106}(NH_3)_{16}(H_3PO_4) + 138O_2$$

This relationship describes the reaction of CO_2 with hydrogen (H), nitrate (NO_3^-), phosphate (H_3PO_4),



Marine Biogeochemistry. Figure 1
Generalized *vertical* distributions of dissolved inorganic elements and particulate matter produced by phytoplankton photosynthesis. Particulate matter concentrations are less than those expected from the disappearance of inorganic elements because of removal by various processes to depth (see Fig. 2). Similarly, the particulate matter *vertical* distribution is less uniform because the time-scale of redistribution of particles is much faster than that of the inorganic elements

and water (H_2O) within the photosynthetic process to produce (↔) organic carbon-nitrogen-phosphorus compounds $\left[(CH_2O)_{106}(NH_3)_{16}H_3PO_4\right]$ and gaseous oxygen (O_2). The numbers preceding the compounds indicate the amount of each. The relationship is reversible (↔) because metabolism (oxidation) of the organic matter produced by photosynthesis regenerates inorganic C, nitrogen (N), and phosphorus (P) in the same ratio and utilizes oxygen. The C:N:P ratio of 106:16:1 obtained from the above relationship is a basic paradigm of marine biogeochemistry. However, Redfield recognized that the C:N:P ratios vary within plankton types and with time, a fact that has been further established in more recent studies. The departure from the basic ratio provides insights in how marine ecosystems change and/or adapt to modified environmental or biological conditions.

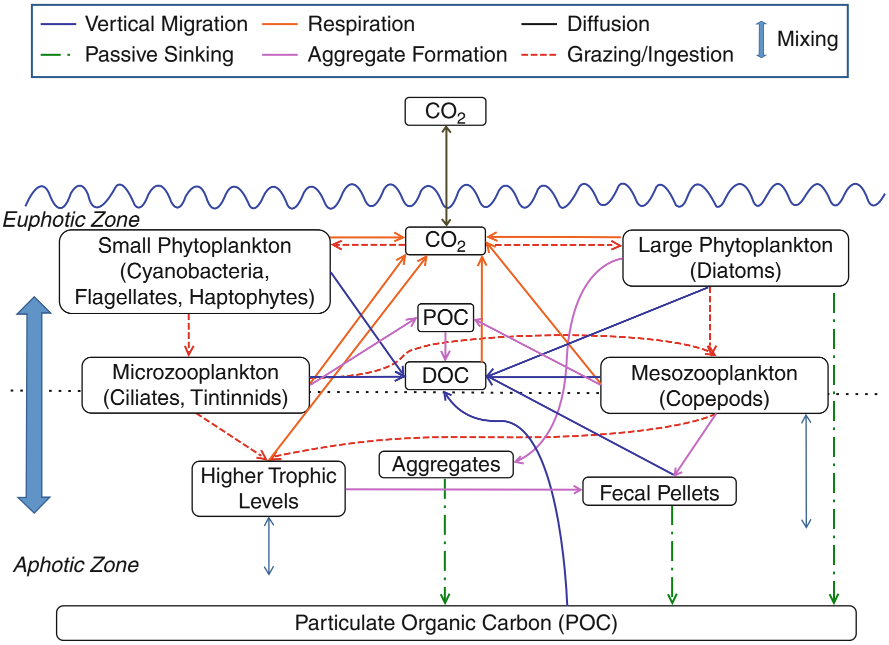
All marine organisms contribute to the carbon cycle by moving carbon between organic and inorganic forms, but some marine organisms are able to use calcification to transform inorganic carbon, using bicarbonate and dissolved calcium from the water column to produce calcium carbonate ($CaCO_3$), which is then used to form a skeleton or protective shell [2]. The dissolution of calcium carbonate back into its original components is one of the primary means by which the particulate components reenter the water column, keeping the inorganic carbon cycle running. Although some of the calcium carbonate dissolved back into the water column comes from dead organisms, a large portion is contributed by phytoplankton from coccolithophorids, the genus coccolithophorid, which produce and shed calcium carbonate shells, making them a major contributor to the inorganic carbon cycle [2]. The calcium carbonate not immediately dissolved back into the water column is removed by sinking, with coccolithophorids comprising a major component of the carbon found in marine sediments. A by-product of calcification is CO_2, which either remains in the water column or reenters the biological pump through photosynthesis [2].

The importance of iron (Fe) and silicic acid ($Si(OH)_4$) in regulating carbon production in oceanic systems has also been established. Iron is required by all living organisms for a variety of metabolic processes, and silicon (Si) is needed by an important phytoplankton functional group, the diatoms, which are characterized by a hard silica shell. Diatoms remove silicic acid in approximately a 1:1 ratio to N, and the P:Fe ratio is approximately 1,000:1. Both ratios show considerable plasticity and their uptake ratios are related to other environmental variables as well [3, 4].

The organic material produced in the upper water column via photosynthesis is used by heterotrophic organisms (e.g., bacteria, zooplankton) and transformed by their metabolism and growth processes. The unassimilated ingestion of these organisms (fecal pellet production) sinks and is oxidized below the euphotic zone by a host of heterotrophic organisms (from bacteria to ciliates to scavenging, mobile animals), thereby converting the organic matter to CO_2. Also, particle aggregates formed from phytoplankton cells, detritus, and dead organisms sink from the euphotic zone and are oxidized. The unidirectional movement of large particles to depth and their remineralization defines the biological pump (Fig. 2), which also contributes to the generation of "nutrient-like" profiles in the ocean. The processes that contribute to the fluxes within the biological pump are critical to understanding the marine carbon cycle.

Atmospheric fluxes of CO_2 into and out of the ocean vary spatially. In general, equatorial waters tend to be large sources of CO_2 (net fluxes are from the ocean to the atmosphere). The equatorial Pacific is a large source because it is the site of large-scale upwelling, a process which brings cold water from depth to the surface. These waters are in turn heated by solar radiation, and because the solubility of CO_2 is strongly temperature dependent (CO_2 is less soluble in warm water), it is lost to the atmosphere. Conversely, polar waters are in general sinks for CO_2. Waters there lose heat to the atmosphere, and thus are able to absorb more CO_2. A topic of intense debate is the possible decrease in carbon flux to the waters of the Southern Ocean resulting from recent increases in wind strength, which may have altered the ocean's ability to remove CO_2 [5]. Such changes potentially would have profound impacts on the global carbon budget. At the present time the ocean is a net sink for atmospheric carbon dioxide, and has sequestered at least 25% of all anthropogenic emissions to date.

Ocean Acidification Recently, great concern has been expressed about the increasing concentrations of CO_2


Vertical Migration
Respiration
Diffusion
Mixing
Passive Sinking
Aggregate Formation
Grazing/Ingestion
CO_2
Euphotic Zone
CO_2
Small Phytoplankton (Cyanobacteria, Flagellates, Haptophytes)
Large Phytoplankton (Diatoms)
POC
Microzooplankton (Ciliates, Tintinnids)
DOC
Mesozooplankton (Copepods)
Aggregates
Higher Trophic Levels
Fecal Pellets
Aphotic Zone
Particulate Organic Carbon (POC)

Marine Biogeochemistry. Figure 2
Schematic of the biological pump showing the biological and chemical components and processes involved in the transformation of carbon dioxide (CO_2) to organic matter, and the subsequent transformation, movement, and oxidation of particulate organic carbon (POC) and dissolved organic carbon (DOC). The CO_2 is absorbed from the atmosphere across the air-ocean interface (*wavy lines*) and is transformed by processes in the euphotic (above *dashed line*) and aphotic zone (below *dashed line*). The migration of zooplankton and higher trophic levels within the water column (*light blue lines*) and unidirectional passive sinking of particles of different sizes to depth (*green dot-dashed line*) redistribute organic material. Processes of grazing/ingestion (*red dashed line*), aggregate formation (*red line*), respiration and CO_2 generation (*orange line*), physical mixing (heavy *blue line*), and solubilization, and DOC generation (*dark blue line*) modify the rate at which POC is exported to depth from the surface waters. The POC pool at depth is generally composed of unidentifiable, small particles, whereas the POC pool in the surface is composed of recognizable biota (bacteria, phytoplankton, zooplankton) and variable amounts of detritus

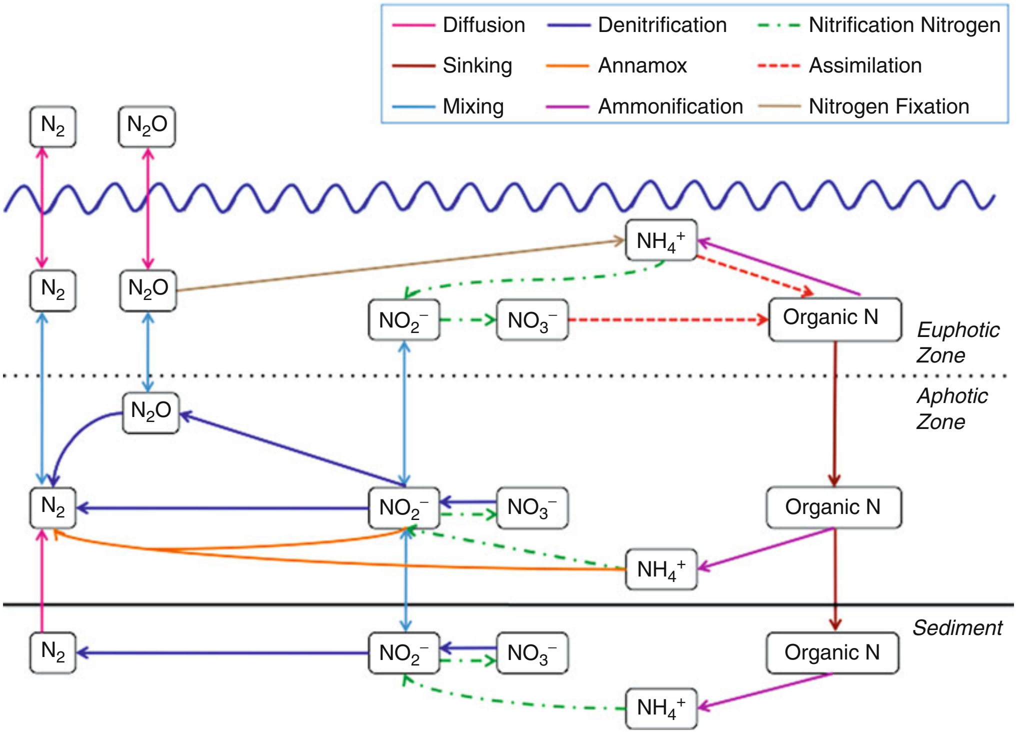
in the ocean, since its absorption decreases the pH, leading to ocean acidification [5, 6]. A decrease in pH would seriously impact calcification, likely increasing dissolution of $CaCO_3$ found in skeletons and shells because the material is unprotected from seawater, and decreasing the rate at which calcification can occur by altering the concentrations of the necessary minerals in the water column. As a result, decreased pH has a great capacity to alter the ecology of marine systems such as coral reefs. In addition, decreased pH levels have been shown to alter the growth, reproduction, efficiency, and survival of those organisms that require $CaCO_3$ to survive, and these effects vary among organisms, suggesting that substantial and unexpected impacts on biodiversity could occur [7].

It is now recognized that many phytoplankton can remove only CO_2 for use in photosynthesis. Under preindustrial pH levels, free CO_2 levels could have been at limiting levels, particularly for conditions that produced high concentrations of algae, because photosynthesis naturally increases the pH level. Decreased pH and increased absolute CO_2 levels arising from current conditions might reduce this limitation. Because there is substantial variability among species of phytoplankton in their response to increased CO_2, planktonic biodiversity is at risk [8]. However, certain algal functional groups, such as nitrogen-fixing cyanobacteria, positively respond to increased CO_2 concentrations by increasing their growth and photosynthesis, whereas others can not. Similarly, at least one species of toxin-producing dinoflagellate demonstrated increased growth and modified elemental ratios under increased CO_2 conditions [9], suggesting the possibility of an enhancement of occurrences of harmful algal blooms in the future. Because the marine carbon cycle is intimately linked with the biogeochemical cycles of nitrogen, phosphorus, silicon, and iron, these interactive effects make it extremely difficult to predict what future decreases in oceanic pH will generate. Oceanographers have recognized that increased inorganic carbon levels can have subtle effects on the biota, and much work is being done to document and quantify these effects.

The Biogeochemical Cycles of Nitrogen and Phosphorus

Although the early work of Redfield [1] clearly differentiated between the sources of nitrogen and phosphorus and the regulation of their turnover, they are linked in nature by the processes operating in the biological pump (Fig. 2). Despite this coupling, as well as their linkage to carbon, there are a number of features that distinguish them.

Nitrogen occurs in three reactive, inorganic forms in the ocean: nitrate (NO_3^-), nitrite (NO_2^-), and ammonium (NH_4^+) and the processes that transform and modify these forms make up the nitrogen cycle (Fig. 3). The nitrogen cycle has five major pathways that result in changes in the availability of nitrogen that can be used by plants. Nitrogen fixation removes gaseous nitrogen from the atmosphere, which is then converted by a series of reactions to forms that can be used for plant growth. In the ocean this process occurs primarily in tropical and semitropical environments, and the major algal species responsible for this transformation is *Trichodesmium*. Denitrification results in



Marine Biogeochemistry. Figure 3

Schematic of the nitrogen cycle showing the nine key mechanisms by which nitrogen moves through the water column, which are denitrification (*dark blue line*), nitrification (*green dash-dot line*), nitrogen assimilation (*red dashed line*), nitrogen fixation (*brown line*), ammonification (*purple line*), annamox (*orange line*), mixing (*light blue line*), diffusion (*pink line*), and sinking (maroon line) the reduction of NO_3^- to gaseous nitrogen, usually mediated by bacteria, and results in the loss of nitrogen available for phytoplankton in oceanic systems. These two processes are the primary means by which the ocean biota controls nitrogen biogeochemistry.

Nitrogen assimilation is the process by which nitrate (NO_3^-) and ammonium (NH_4^+) are removed from the water by phytoplankton. Ammonium is energetically favored for uptake because it does not have to be reduced intracellularly, but nitrate often occurs in greater concentrations, particularly in areas of upwelling or deep vertical mixing. Ammonium inhibits nitrate uptake, but the degree of inhibition varies with the relative concentration of the two nutrients. Ammonification generates NH_4^+ by the cleaving of amine groups from organic nitrogen. Because many marine organisms excrete ammonium, the vertical distribution of NH_4^+ can depend on the distribution of heterotrophs, such as copepods, which is variable. Nitrification is the production of NO_3^- from ammonium. Earlier work suggested that this was a relatively slow process, but more recent investigations suggest that the oxidation of NH_4^+ and production of nitrate is quite rapid, particularly in tropical waters.

The different transformations result from different organisms and some require specific types of environmental conditions (Table 1). Denitrification and nitrogen fixation are anaerobic processes, which occur only in the absence of oxygen. Oceanic systems, ranging from estuarine to open ocean, provide sites for denitrification and as a result are depleted of oxygen. These oxygen-minimum regions are characterized by large vertical fluxes of organic matter, which heterotrophic bacteria oxidize and release nitrogen, consuming the available oxygen in the process. Anoxic conditions also occur in sediments where oxygen is depleted by aerobic metabolism. An unusual biological adaptation allows for nitrogen fixation (an anaerobic process) to occur in surface waters with high levels of oxygen. Some organisms (e.g., the cyanobacterium *Trichodesmium)* form extensive patches or tufts. These tufts, by virtue of their own metabolism, unusually thick cell walls and biochemical modifications of specialized cells where N_2 fixation occurs, create a microzone of very low oxygen, thus allowing nitrogen fixation to proceed. Other, smaller cyanobacteria have unusual biochemical adaptations that allow them to fix N_2 as well, despite living in oxygen-saturated water.

Recently a new nitrogen transformation, the annamox pathway, has been described in which anaerobic bacteria oxidize ammonium and nitrite directly to gaseous nitrogen, providing a second means by which nitrogen is "lost" from the nitrogen cycle [10]. This pathway has been found to be quantitatively important in regions such as the Peruvian and Arabian Sea oxygen-minimum zones [11, 12]. Because 30–50% of global nitrogen "losses" occur in these types of regions, elucidation of this process, its oceanographic controls, and the absolute rates, have important implications for the global nitrogen cycle.

Marine Biogeochemistry. Table 1 Summary of the major processes in the nitrogen cycle, the organisms responsible for the different processes, and the environmental conditions necessary for each process

Process	Organism(s) responsible	Necessary environmental condition
Nitrogen fixation [N_2(gas)→ reduced N]	Cyanobacteria, nitrogen-fixing bacteria	Absence of O_2; light for cyanobacteria
Denitrification [$NO_3^- \rightarrow N_2(\mathrm{gas})$]	Denitrifying bacteria	Absence of O_2
Ammonification [$N_{org} \rightarrow NH_4^+$]	Heterotrophic organisms	Presence of O_2
Nitrogen assimilation [$NO_3^- \rightarrow N_{org}; NH_4^+ \rightarrow N_{org}$]	Large phytoplankton/diatoms for NO_3^- uptake; small phytoplankton for NH_4^+ uptake	Light
Nitrification [$NH_4^+ \rightarrow NO_3^-$]	Bacteria	Presence of O_2
Annamox [$NH_4^+, NO_2^- \rightarrow N_2(gas)$]	Bacteria	Absence of O_2

The Biogeochemical Cycle of Iron

The understanding of the role of iron in the ocean has undergone a dramatic revision in the past few decades. Until recently data on absolute iron concentrations were seriously compromised by the difficulty of obtaining samples without contamination. As the collection and sampling aspects were greatly improved, the ability to quantify concentrations of iron in the oxygenated waters of the ocean decreased dramatically. Coincident with increased realization and acceptance of the vanishingly low concentrations of iron was the hypothesis that iron could, and does, regulate phytoplankton growth and productivity over large areas of the ocean [13]. Indeed, the hypothesis appeared to explain a number of oceanic features that were only partially explained. For example, large areas of the ocean, such as the Southern Ocean, the equatorial Pacific, and the north Pacific, have substantial standing stocks of nitrate and phosphate, as well as adequate irradiance, but exhibit very low standing stocks of phytoplankton (high-nutrient, low-chlorophyll regions, or HNLCs). During glacial-interglacial periods, atmospheric concentrations of CO_2 showed substantial variations and were strongly negatively correlated with iron deposition [14]. Thus iron limitation could explain CO_2 variations over geological time as well. Given that iron is the fourth most abundant element on earth, how can such low concentrations exist in the ocean, and how did oceanographers unequivocally demonstrate the ecological importance of iron?

Iron is derived from terrestrial and hydrothermal sources, but upon entry into oxygenated, saline waters, it rapidly forms iron oxides. The precipitates are largely insoluble under aerobic conditions, and attach to particles or remain in the water as colloids. The colloids can be solubilized by irradiance, contributing to a pool of dissolved inorganic iron, which consists of two forms, Fe^{+2} and Fe^{+3}. Both of these ions can be removed by plankton for their growth, although Fe^{+2} is generally oxidized to Fe^{+3} and kept at low levels. The mean ocean concentration of dissolved inorganic iron in the upper 200 m of the ocean is 0.07 nmol kg^{-1} [12]. Both forms can also be chelated by organic molecules, and thus become part of the dissolved ferro-organic pool. In general, there are two classes of organic ligands that bind with iron, a strong-binding ligand and a weak-binding ligand. The latter exchanges iron easily with biota, and thus makes iron bioavailable. There is also a class of special ligands called siderophores, which are low molecular weight organics that are produced and excreted primarily by prokaryotic organisms (bacteria, cyanobacteria) and that bind dissolved inorganic iron [15]. The ferro-ligand complex can be assimilated by bacteria, phytoplankton, and cyanobacteria, and the iron incorporated into a variety of cellular processes. Transformations among all of these pools are both biologically and irradiance mediated; entirely different transformations and equilibria are established in anoxic waters and sediments.

Iron in ocean surface waters derives from either atmospheric or deep ocean sources. Atmospheric deposition varies by latitude (proximity to terrestrial sources) and temporally (dependent on source region wind variability). Aerosols can be measured by satellite-borne sensors, which have shown that some oceanic systems receive substantial periodic depositions of iron from industrial sources (the North Atlantic) and from dust derived from terrestrial deserts in China (the western Pacific) and the Sahara in Africa (the coast of North Africa). Dissolution of aerosols in ocean water (fractional solubility) depends on the type of mineral in the aerosol, and can range from <1–90% [16, 17]. Small aerosol particles can rapidly aggregate with biological particles and exit the surface layer by sinking. Residence times for particulate iron can be as short as 6 days [17]. Conversely, other regions are rarely impacted by atmospheric deposition events (e.g., the Southern Ocean, the equatorial Pacific) by virtue of large-scale wind patterns that isolate them from terrestrial sources. These regions have their iron inputs driven by oceanographic processes such as deep vertical mixing and upwelling. Given the spatial and temporal variability in both of these processes, it is not surprising that surface water concentrations of iron are also highly variable.

Mesoscale Iron Fertilization Experiments

In the 1990s a series of large-scale ocean manipulations were undertaken to test the hypothesis that iron limited phytoplankton growth in the tropical Pacific. Two competing hypotheses were offered to explain the equilibrium concentrations of high concentrations of nitrate and low phytoplankton biomass which were (1) limiting levels of bioavailable iron and (2) rates of loss processes from grazing kept phytoplankton standing stocks at low levels. To test these, in situ additions of iron were planned for limited regions of the ocean. The passive tracer sulfur hexafluoride (SF_6), which can be detected at very low levels, was added with the iron so that the enriched water could be followed over time. The first iron enrichment experiment produced contradictory results. The photosynthetic capacity of phytoplankton showed a clear enhancement that was correlated with iron additions, but nitrate and CO_2 concentrations were unaffected [18, 19]. Further analysis showed that upon initial iron enrichment, the iron dropped to extremely low levels because colloid formation rapidly converted soluble iron to insoluble iron oxides, and the fertilized water patch was subducted to depth, which removed the iron-enriched water from the high irradiance euphotic zone required for nutrient assimilation. To further test the two hypotheses, the experiment was repeated, and this experiment clearly demonstrated the critical role of iron in limiting phytoplankton growth in high-nutrient, low-chlorophyll waters. Iron was added repeatedly to the patch of water at 3-day intervals for almost 2 weeks [20], and the response of the surface water was clear, showing decreased nitrate (which dropped to zero), decreased CO_2, increased phytoplankton biomass and photosynthetic activity, and a quantifiable decrease in iron concentrations. That is, the concentration and supply of iron was nevertheless the essential feature in driving the carbon and nitrogen cycles of the equatorial Pacific Ocean.

Subsequent similar iron enrichment experiments have been conducted in other HNLC regions in the Southern Ocean and the North Pacific. The former is extremely important to global biogeochemical cycles, as it is the site of deep and intermediate water mass formation, and thus regulates the concentrations of inorganic nutrients in much of the world's surface waters. As an example, models suggested that if all the inorganic nutrients were utilized (by iron fertilization) in the Southern Ocean that within 300 years the waters being upwelled in the eastern tropical Pacific would be greatly reduced in nutrient levels, and thus decrease productivity of commercially important higher trophic levels and marine mammals dependent on ecosystem processes in that region [21]. In all iron enrichment experiments to date, substantial and positive responses to additions of inorganic iron were observed, and while the details among experiments differ (and the causes debated), it is now accepted that iron plays a major role in the biogeochemistry of the ocean [22].

The Biogeochemical Cycle of Silicon

Silicon, despite being a nutrient for only one major functional group of phytoplankton (diatoms), is a major factor in regulating other biogeochemical cycles, such as carbon. This is because diatoms are extremely important primary producers, generating approximately as much oxygen on an annual basis as do pine trees in terrestrial systems. In addition, diatoms are among the largest forms of phytoplankton, and hence can sink passively to depth. Diatoms also produce transparent exopolymer particles, which serve as the primary mechanism for aggregating particles in the ocean's surface layer, thus producing large, rapidly sinking particles that are the major component of organic carbon and nitrogen flux to deeper water (Fig. 2). Finally, diatoms are also heavily grazed by herbivorous organisms, and serve as a means to transfer photosynthate to the large organism-based food web. All of these characteristics contribute to the substantial importance of diatoms in the ocean.

Silicon is a major component of rocks and terrestrial minerals, and as a result the inputs to the ocean in riverine waters are substantial. However, silicon is not readily dissolvable, and dissolved silicon, which occurs as $Si(OH)_4$, remains at relatively low levels. Aeolian and oceanic weathering of seafloor rocks also constitutes a significant source of dissolved silicon. Silicon also is found in high concentrations in waters exiting hydrothermal vents, and while quantitative estimates are uncertain, the contribution of this source to total silicon inputs is likely to be significant.

Silicon is incorporated into diatoms and other marine organisms as opal $(Si(OH)_4 nH_2O)$, which is slightly more soluble than pure SiO_2 and undersaturated in all ocean waters. Opal is found in the sediments as siliceous deposits of biogenic origin; these deposits are largely focused in the Southern Ocean's polar front region [23]. Silicon is recycled within the water column, but rates of this cycling are modest, and silicon regeneration is often markedly uncoupled from that of carbon and nitrogen in some regions. The reason for this appears to result from the different controls of each: organic matter regeneration is largely biologically mediated (by heterotrophic processes), whereas silicon regeneration is regulated by temperature [24]. As a result, in polar regions a large fraction of the organic matter that sinks from the euphotic zone is regenerated in the upper 250 m, whereas a substantial amount of silicon sinks to a greater depth as biogenic particles. This uncoupling contributes to the formation of large zones of biogenic silica deposits in polar regions and are reflective of surface layer diatomaceous productivity. In a more recent reanalysis of the global silicon budget, it was concluded that the deposition of silicon in continental margins may have been greatly underestimated [25]. If this were true, then the coupling between the silicon and organic matter budgets would be even stronger than previously thought.

An additional mechanism to couple the biogeochemistry of silicon and organic carbon is the presence of an organic membrane that covers diatom frustules [26]. Silica dissolution does not begin until this membrane is degraded by bacteria, which decreases the time for the dissolution of opal during the transit of a particle through the water column (ca. 3,000 m). Sinking rates of large aggregates are ca. 200 m day^{-1}, so that a reduction in the already low rate of dissolution by the necessity for organic degradation can decrease dissolution of silica markedly. Similar effects of grazing can occur, as fecal pellets are usually composed of an organic pellicle that must be degraded prior to chemical silica dissolution.

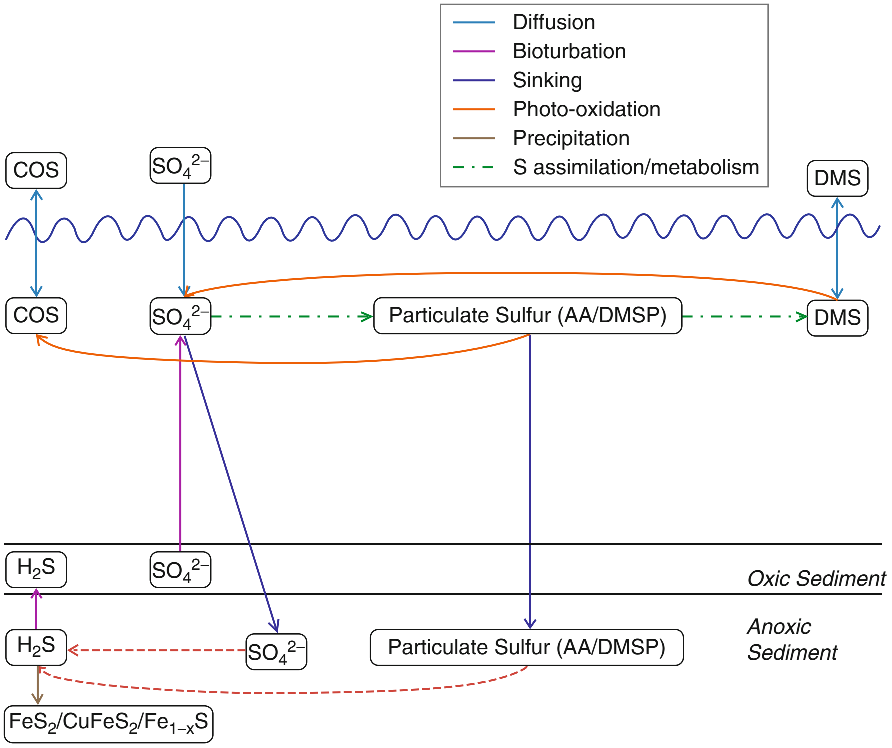
As with other nutrients, silicic acid has substantial interactions with other elements, such as nitrate and iron. Under iron-limiting conditions, diatoms continue to assimilate silicon, but because iron is needed in the enzymes used for nitrate assimilation, nitrate uptake decreases [3]. As a result, Si:N ratios increase by nearly an order of magnitude in diatoms under iron limitation and elevated ratios observed in natural systems have been used to infer iron limitation.

The Biogeochemical Cycle of Sulfur

In marine systems sulfur is largely present in its most stable form, which is sulfate ($SO_4{}^{-2}$). Sulfate is present in high concentrations in most marine systems, and relatively low concentrations are required by organisms to survive [27]. As a result, sulfur does not normally become growth limiting. Sulfate concentrations in marine systems are primarily controlled by physical rather than chemical processes. Variations in concentration only have a significant biological impact in anoxic zones where sulfate reduction occurs [2]. Sulfur is also present as other inorganic (H_2S) and organic (dimethylsulfoniopropionate (DMSP), dimethylsulfide (DMS), carbonyl sulfide (COS), and methanethiol (MeSH)) forms. Sulfate is transformed into these compounds via the sulfur cycle, which operates primarily in the photic zone of the upper water column, in the sediments, and around hydrothermal vents (Fig. 4).

In aerobic environments sulfur is converted between inorganic compounds (sulfate and hydrogen sulfide) and organic sulfur compounds including DMSP, DMS, COS, and amino acids. Most algae and bacteria use sulfur assimilation to form amino acids, such as cysteine and methionine [27]. Some phytoplankton species, particularly prymnesiophytes and dinoflagellates, use methionine to produce DMSP, a compound with antioxidant properties [28, 29]. DMSP can be released into the water and subsequently used to produce amino acids through assimilation by bacteria or phytoplankton, including some species of diatoms and cyanobacteria, demethylated by bacteria to produce MeSH, or oxidized into DMS and acrylic acid [30]. DMS is either broken down in the water into sulfate through bacterial uptake or photooxidation, or is volatilized into the atmosphere, where it can act as an important aerosol [27].

The sulfur cycle in ocean sediments can be divided into reactions that occur in the upper oxic layer and those that occur in the lower, oxygen-depleted (anoxic) region. In the anoxic sediments, sulfur-reducing bacteria carry out anaerobic respiration using sulfate or sulfur-containing organic compounds to oxidize organic matter, resulting in the production of sulfide, typically as H_2S, a form of sulfur that is highly toxic to most organisms. In the deeper layers of the sediment, sulfide reacts with iron and precipitates as iron sulfides such as pyrite (FeS_2) [30]. Some sulfide remains in the sediment, and, when mixed back into the oxic zone through processes such as bioturbation, is quickly



Marine Biogeochemistry. Figure 4

The sulfur cycle involves transformation of sulfur in the water column through physical mechanisms such as diffusion (*light blue line*), bioturbation (*purple line*), and sinking (*dark blue line*); chemical mechanisms of photooxidation (*orange line*), and precipitation (*yellow line*); and biological mechanisms of sulfur assimilation and metabolism by phytoplankton (*green dash-dot line*) and reduction and oxidation by bacteria in the sediment (*orange dashed line*)

oxidized by sulfur-oxidizing bacteria into sulfate, which can then remain in the sediment or be released into the overlying water [31]. Sulfur oxidation and reduction by bacteria in the sediment are also important to the functioning of the nitrogen cycle in oxygen-minimum zones [32]. In these environments, sulfate reduction provides a significant amount of the ammonium used in the anammox reaction in anaerobic environments, and nitrate reduction may be coupled to sulfide oxidation, indicating that the anaerobic mechanisms in the sulfur cycle may also be important in the nitrogen cycle [32].

The presence of hydrogen sulfide around hydrothermal vents has resulted in the development of unique organisms with the ability to use the energy contained in hydrothermal fluids to produce organic compounds through chemoautolithotrophy [33]. At hydrothermal vents seawater comes into contact with magma from the earth's interior, which cools and forms reduced sulfur compounds [2]. The sulfate in seawater then reacts to form hydrogen sulfide as well as sulfur-containing minerals such as pyrite (FeS_2), chalcopyrite ($CuFeS_2$), and pyrrhotite ($Fe_{1-x}S$) [2], which form the surface chimney structure that is characteristic of hydrothermal vents. The hydrogen sulfide provides the energy, rather than light, for the chemoautrophic microorganisms that form the base of the hydrothermal vent food web [33]. Some species of microorganisms can operate in aerobic conditions, using oxygen as the electron acceptor, while others have the ability to carry out this reaction in anaerobic conditions, using nitrate, sulfate, or sulfur as the electron acceptor [2]. These organisms survive in symbiotic relationships with other organisms living near the hydrothermal vents. The microorganisms, which are endemic to hydrothermal vent environments, allow unique communities to develop and help maintain the oceanic sulfur cycle by transforming hydrogen sulfide released by the hydrothermal vents into sulfate [33].

The Biogeochemical Cycle of Oxygen

Oxygen is involved in all nutrient cycles, and its presence or absence dictates the reactions that will occur in a specific marine environment. Oxygen gas can be introduced into marine environments across the air-sea interface (e.g., by diffusion). However, oxygen concentration is controlled by the biological processes of photosynthesis and respiration, and by physical processes such as mixing within the water column. In the euphotic zone, phytoplankton photosynthesis produces oxygen, which is then used as the electron acceptor to conduct aerobic respiration. This process is carried out by both autotrophic and heterotrophic organisms throughout the water column.

Oxygen concentration generally decreases with depth in the ocean. Photosynthesis can only be carried out in the lighted parts of the water column, but respiration continues throughout the water column. As the organic matter from the surface layers sinks, it is taken up by organisms and used to conduct respiration, depleting oxygen levels. Some marine environments, particularly in marine sediments, are suboxic, with oxygen concentrations less than 0.2 ppm (but still detectable), or anoxic, with oxygen concentrations below detectable levels [2]. Organisms survive in these environments by using anaerobic respiration, in which compounds such a nitrate, sulfate, iron, or even organic matter are used as alternative electron acceptors to oxygen [2].

Anoxic zones are not limited to marine sediments, with increasing attention being paid to decreasing oxygen concentrations in previously oxygen-rich areas of the ocean. Hypoxic zones, marine environments with oxygen concentrations below 2 mg L^{-1}, typically form when primary productivity is high, leading to increased organic matter in the system and increased respiration, and when mixing throughout the water column is low, preventing the oxygen in the upper water column from reaching lower layers [34–36]. Hypoxic zones have been increasing in frequency, including the Gulf of Mexico and Chesapeake Bay [34, 35]. Factors such as eutrophication due to increased fertilizer or wastewater runoff have lead to the development of hypoxic conditions in systems already susceptible due to vertical stratification of the water column [34]. Thus, the disruption in the typical oxygen cycle and the lack of an anaerobic respiration mechanism in most marine organisms can result in serious consequences for the composition and productivity of the marine food web community in these hypoxic zones.

Future Directions

Studies of Biogeochemical Cycles

In the past two decades, a number of large, interdisciplinary programs were conducted to obtain biogeochemical data on appropriate time and space scales so that mathematical models of global climate change can accurately represent the complex processes of elemental cycles. One such program, the Joint Global Ocean Flux Study (JGOFS), which occurred from 1987 to 2003, was international in scope, and undertook coordinated, multidisciplinary, international studies in the equatorial Pacific, the north Atlantic, the Arabian Sea, and the Southern Ocean, and coordinated multidisciplinary national programs in a range of coastal and open ocean environments. The JGOFS project was designed to assess the carbon cycle, but because all elemental cycles are closely linked, insights were gained into the understanding of nitrogen, silicon, and iron cycles as well. The JGOFS program also had a significant synthesis and modeling component that was intended to integrate the data sets from the multidisciplinary studies and to develop mathematical models of increased complexity and biological realism. In addition to providing a wealth of publicly available data, the JGOFS program served as a model for large, multidisciplinary studies of ocean processes.

The results and understanding from the JGOFS program provided the basis for the Integrated Marine Biogeochemistry and Ecosystem Research (IMBER) Project, which was initiated in 2001 by the International Geosphere-Biosphere Program and the Scientific Committee on Oceanic Research. The science goals of the IMBER project extend the investigation of marine biogeochemical cycles to include the influence of feedbacks with marine food webs and the consequences for marine ecosystems. Central to the IMBER goal is the development of a predictive understanding of how marine biogeochemical cycles and ecosystems respond to complex forcings, such as large-scale climatic variations, changing physical dynamics, carbon cycle chemistry and nutrient fluxes, and the impacts of marine harvesting. IMBER science is making new advances in understanding marine systems by bringing together the natural and social science communities to study key impacts and feedbacks between the marine and human systems. The emerging recognition of human interactions as integral parts of marine ecosystems is providing the direction for future integrative research designed to understand and sustain ocean systems as environmental change and its associated uncertainties occur.

Role of Modeling

Mathematical models provide an approach for integrating and synthesizing the knowledge and understanding obtained from measurements of oceanic biogeochemical processes. The use of biogeochemical models in ocean research has a long history [37, 38] but their use was advanced significantly in the early 1990s when a model that simulated nitrogen cycling through the lower trophic levels in the oceanic mixed layer became generally available [39], which subsequently has provided the basis for the coupled circulation-biogeochemical models that are now embedded in regional, basin, and global scale models.

The skill of the current generation of biogeochemical models is sufficient to allow projections of future states that may result from climate variability and the oceanic uptake of anthropogenic carbon [40–42]. The patterns and distributions emerging from these simulations show shifts in phytoplankton distributions and marine biomes, alteration of phytoplankton species assemblages, and modified lower trophic level community structure [43–45], all of which have direct and important consequences for biogeochemical cycling. Simulations of the effect of increasing atmospheric CO_2 and its uptake by the ocean show reductions in ocean pH and in saturation levels of calcium carbonate, which have serious consequences for many marine organisms [46].

Advances in conceptual understanding, modeling techniques, and data availability have made predictive marine biogeochemical models a feasible goal [47]. However, modeling for prediction is still rapidly developing and much remains to be done in generating appropriate frameworks and in collection of data sets that support predictive modeling for marine biogeochemical cycling [48, 49].

Bibliography

Primary Literature

1. Redfield AC (1958) The biological control of chemical factors in the environment. Am Sci 64:205–221
2. Libes SM (2009) Introduction to marine biogeochemistry. Academic, Amsterdam
3. Hutchins DA, Bruland KW (1998) Iron-limited diatom growth and Si:N uptake ratios in a coastal upwelling regime. Nature 393:561–564
4. Sunda WG, Huntsman S (1997) Interrelated influence of iron, light, and cell size on growth of marine phytoplankton. Nature 390:389–392
5. Le Quéré C et al (2007) Saturation of the Southern Ocean CO_2 sink due to recent climate change. Science 316:1735–1738
6. Doney SC, Fabry VF, Feely RA, Kleypas JA (2009) Ocean acidification: the other CO_2 problem. Ann Rev Mar Sci 1:169–192
7. Ries BR, Cohen AL, McCorkle DC (2009) Marine calcifiers exhibit mixed responses to CO_2-induced ocean acidification. Geology 37:131–134
8. Hutchins DA, Mulholland MR, Fu F (2009) Nutrient cycles and marine microbes in a CO_2-enriched ocean. Oceanogr 22:128–145
9. Fu F-X, Zhang Y, Warner ME, Feng Y, Hutchins DA (2008) A comparison of future increased CO_2 and temperature effects on sympatric *Heterosigma akashiwo* and *Prorocentrum minimum*. Harm Algae 7:76–90
10. Devol AH (2003) Solution to a marine mystery. Nature 422:575–576
11. Lam P et al (2009) Revising the nitrogen cycle in the Peruvian oxygen minimum zone. Proc Nat Aca Sci 106:4752–4757
12. Ward B et al (2009) Denitrification as the dominant nitrogen loss process in the Arabian Sea. Nature 451:78–81
13. Martin JH, Fitzwater S (1988) Iron deficiency limits phytoplankton growth in the northeast Pacific subarctic. Nature 331:341–343
14. Ducklow HW, Oliver JL, Smith WO (2007) The role of iron as a limiting nutrient for marine plankton processes. In: Melillo JM, Field CB, Moldan B (eds) Interactions of the major biogeochemical cycles. Island Press, Washington, DC, pp 295–310
15. Vraspir JM, Butler A (2009) Chemistry of marine ligands and siderophores. Ann Rev Mar Sci 1:43–63
16. Johnson KS, Gordon RM, Coale KH (1997) What controls dissolved iron concentrations in the world ocean? Mar Chem 57:137–161
17. Baker AR, Croot PR (2010) Atmospheric and marine controls on aerosol iron solubility in seawater. Mar Chem 120:4–13
18. Martin JH et al (1994) Testing the iron hypothesis in ecosystems of the equatorial Pacific Ocean. Nature 371:123–129

19. Watson AJ et al (1994) Minimal effect of iron fertilization on sea-surface carbon dioxide concentrations. Nature 371:143–145
20. Coale KH et al (1996) A massive phytoplankton bloom induced by an ecosystem-scale iron fertilization experiment in the equatorial Pacific. Nature 383:495–501
21. Sarmiento JL, Gruber N, Brzezinski MA, Dunne JP (2004) High-latitude controls of thermocline nutrients and low latitude biological productivity. Nature 427:56–60
22. Boyd PW et al (2007) Mesoscale iron enrichment experiments 1993–2005: synthesis and future directions. Science 315:612–617
23. Treguer P, Nelson DM, van Bennekom AJT, DeMaster DJ, Lynaert A, Queguiner B (1995) The silica balance in the world ocean: a reestimate. Science 268:375–379
24. Nelson DM et al (1995) Production and dissolution of biogenic silica in the ocean: revised global estimates, comparison with regional data and relationship to biogenic sedimentations. Global Biogeochem Cycles 9:359–372
25. DeMaster DJ (2002) The accumulation and cycling of biogenic silica in the Southern Ocean: revisiting the marine silica budget. Deep Sea Res II 49:3155–3167
26. Biddle KD, Azam F (1999) Accelerated dissolution of diatom silica by natural bacterial assemblages. Nature 397:508–512
27. Sievert SM, Kiene RP, Schulz-Vogt HN (2007) The sulfur cycle. Oceanography 20:117–123
28. Andreae MO (1990) Ocean-atmosphere interactions in the global biogeochemical sulfur cycle. Mar Chem 30:1–29
29. Sunda W, Kieber DJ, Kiene RP, Huntsman S (2002) An antioxidant function for DMSP and DMS in marine algae. Nature 418:317–320
30. Jorgensen BB (1977) The sulfur cycle of a coastal marine sediment (Limfjorden, Denmark). Limnol Oceanogr 22:814–832
31. Canfield DE, Farquhar J (2009) Animal evolution, bioturbation, and the sulfate concentration of the oceans. Proc Nat Acad Sci 106:8123–8127
32. Canfield DE, Stewart FJ, Thamdrup B, Brabandere LD, Dalsgaard T, Delong EF, Revsbech NP, Ulloa O (2010) A cryptic sulfur cycle in oxygen-minimum-zone waters off the Chilean coast. Science 330:1375–1378
33. van Dover CL (2000) The ecology of deep-sea hydrothermal vents. Princeton University Press, Princeton
34. Diaz RJ, Rosenberg R (2008) Spreading dead zones and consequences for marine ecosystems. Science 321:926–929
35. Scavia D, Justic D, Bierman VJ (2004) Reducing hypoxia in the Gulf of Mexico: advice from three models. Estuaries 27:419–425
36. Lam P, Kuypers MMM (2011) Microbial nitrogen cycling processes in oxygen minimum zones. Ann Rev Mar Sci 3:317–345
37. Riley GA (1946) Factors controlling phytoplankton populations on Georges Bank. J Mar Res 6:54–73
38. Riley GA (1947) A theoretical analysis of the zooplankton populations of Georges Bank. J Mar Res 6:104–113
39. Fasham MJR, Ducklow HW, McKelvie SM (1990) A nitrogen-based model of plankton dynamics in the oceanic mixed layer. J Mar Res 48:591–639
40. Orr JC et al (2001) Estimates of anthropogenic carbon uptake from four three-dimensional global ocean models. Global Biogeochem Cycles 15:43–60
41. Doney SC et al (2004) Evaluating global ocean carbon models: the importance of realistic physics. Global Biogeochem Cycles 18:GB3017. doi:10.1029/2003GB002150
42. Boyd PW, Doney SC (2001) Modeling regional response by marine pelagic ecosystems to global climate change. Geophys Res Lett 29. doi:10.10292001GL014130
43. Follows MJ, Dutkiewicz S, Grant S, Chisholm SW (2007) Emergent biogeography of microbial communities in a model ocean. Science 315:1843–1846
44. Le Quéré C et al (2005) Ecosystem dynamics based on plankton function types for global ocean biogeochemistry models. Global Change Biol 11:2016–2040. doi:10.111/j.1365-2486.2005.1004.x
45. Sinha B, Buitenhuis ET, Le Quéré C, Anderson TR (2010) Comparison of the emergent behavior of a complex ecosystem model in two ocean general circulation models. Prog Oceanogr 84:204–224
46. Orr JC et al (2005) Anthropogenic ocean acidification over the twenty-first century and its impact on calcifying organisms. Nature 437:681–686
47. Friedrichs MAM et al (2007) Assessment of skill and portability in regional marine biogeochemical models: role of multiple planktonic groups. J Geophys Res 112:C08001. doi:10.1029/2006JC003852
48. Anderson TR (2009) Progress in marine ecosystem modeling and the “unreasonable effectiveness of mathematics.” J Mar Sys 81:4–11
49. Ducklow HW, Doney SC, Steinberg DK (2009) Contributions of long-term research and time-series observation to marine ecology and biogeochemistry. Ann Rev Mar Sci 1:279–302

Books and Reviews

Baliño BM, Fasham MJR, Bowles MC (2001) Ocean biogeochemistry and global change. IGBP Science no 2, Stockholm (A popular summary of JGOFS main achievements)

Benitez-Nelson CR (2000) The biogeochemical cycling of phosphorus in marine systems. Earth-Sci Rev 51:109–135

Brandes JA, Devol AH, Deutsch C (2007) New developments in the marine nitrogen cycle. Chem Rev 107:577–589

Broecker WS (2009) Wally’s quest to understand the ocean’s $CaCO_3$ cycle. Ann Rev Mar Sci 1:1–18

Deutsch C, Brix H, Ito T, Frenzel H, Thompson L (2011) Climate-forced variability of ocean hypoxia. Science 333:336–339

Diaz RJ (2000) Overview of hypoxia around the world. J Environ Quality 30:275–281

Doney SC et al (2001) Marine biogeochemical modeling: recent advances and future challenges. Oceanography 14:93–108

Fasham MJR (ed) (2003) The role of the ocean carbon cycle in global change. Springer, Heidelberg, p 297

Gruber N, Galloway JN (2008) An earth-system perspective of the global nitrogen cycle. Nature 451:293–296

Hanson RB, Ducklow HW, Field JG (eds) (2000) The changing ocean carbon cycle: a midterm synthesis of the joint global ocean flux study, IGBP book series no 5. Cambridge University Press, Cambridge, 520 pp

IMBER Science Plan and Implementation Strategy (2005) IGBP Report N°52. IGBP Secretariat, Stockholm, 76 pp

Jickells TD et al (2005) Global iron connections between desert dust, ocean biogeochemistry, and climate. Science 308:67–71

Keeling RF, Kortzinger A, Gruber N (2010) Oxygen deoxygenation in a warming world. Ann Rev Mar Sci 2:199–229

Paulmier A, Ruiz-Pino D (2009) Oxygen minimum zones (OMZs) in the modern ocean. Prog Oceanogr 80:113–128

Ragueneau O et al (2000) A review of the Si cycle in the modern ocean: recent progress and missing gaps in the application of biogenic opal as a paleoproductivity proxy. Global Planetary Change 26:317–365

Sabine CL, Tanhua T (2010) Estimation of anthropogenic CO_2 inventories in the ocean. Ann Rev Mar Sci 2:175–198

Sunda WG (2010) Iron and the carbon pump. Science 327:654–655

Marine Fisheries Enhancement, Coming of Age in the New Millennium

Kenneth M. Leber
Center for Fisheries Enhancement, Mote Marine Laboratory, Sarasota, FL, USA

Article Outline

Glossary

Anadromous Species that spawn in freshwater, then their offspring gradually make their way into estuaries or the sea, where they remain during much of the subadult and adult stages of the life cycle, before returning to rivers and streams to spawn.

Catadromous Species whose females release their eggs at sea, then the offspring move as larvae or early juveniles into estuaries, rivers, and streams where they spend the juvenile stage of the life cycle.

Marine Species that spawn in sea water, including those that spend most of their lives at sea and catadromous fishes, which spawn in seawater, then enter freshwater nursery habitats.

Marine fisheries enhancement Release of aquacultured marine organisms into seas and estuaries to increase or restore abundance and fishery yields in the wild.

Outbreeding depression Caused when offspring from crosses between individuals from different populations or subpopulations (stocks) have lower fitness than progeny from crosses between individuals from the same population/stock.

Recruitment The process of joining an existing population. Species *recruit* to the juvenile stages in nursery habitats; juveniles subsequently *recruit* to adult stages in adult habitats. Species *recruit* to a fishery when they reach the minimum size fished.

Reintroduction Temporary release of cultured organisms with the aim of reestablishing a locally extinct population.

Restocking Release of cultured juveniles into wild population(s) to restore severely depleted spawning biomass to a level where it can once again provide regular, substantial yields.

Sea ranching Release of cultured juveniles into unenclosed marine and estuarine environments for harvest at a larger size in “put, grow, and take” operations.

Stock enhancement The release of cultured juveniles into wild populations to augment the natural supply of juveniles and optimize harvests by overcoming limitations in juvenile recruitment.

Supplementation Moderate release of cultured fish into very small and declining populations, with the aim of reducing extinction risk and conserving genetic diversity. Supplementation serves primarily conservation aims and specifically addresses sustainability issues and genetic threats in small and declining populations.

Definition of the Subject

Marine fisheries enhancement (aka "stock enhancement") is the use of hatchery-reared saltwater organisms to increase abundance and fishery yields in the wild. "Conservation hatcheries" also produce and stock depleted, threatened, or endangered organisms – to help preserve species in decline. The practice began in the latter part of the nineteenth century when fish hatcheries were first developed but understanding of the ecology and management of wild stocks into which the hatchery-reared organisms where released was very limited. Early stock enhancement thus has gone through a series of fits and starts and misfires. In the century after its birth, the technologies required for scientific inquiry of the effects and effectiveness of stocking hatchery-reared organisms were lacking. The science needed to guide reliable use of cultured aquatic organisms in conservation and resource management remained undeveloped. Then, at the close of the twentieth century, new mariculture, tagging, and genetic technologies surfaced and rapid advances were made in the science underpinning marine stock enhancement.

As growth in human population size approaches the carrying capacity of the planet in this century, and the world increasingly turns to the oceans to farm and harvest food [1], sustainable fishery yields and conservation of natural resources face unparalleled challenges. Over the past two decades, marine fisheries enhancement has been transformed from a tentative, poorly developed management tool to a maturing science. Some believe research funding for this field would be better spent on traditional fishery management. But today's seafood producers, fishery managers, and "…conservationists need all the tools that biology, ecology, diplomacy and politics can muster if endangered species are to survive beyond the next century," [2] and fisheries are to continue to support a viable seafood industry and sport pastime. This entry traces the emergence and progress of marine fisheries enhancement, and offers a prescription for future direction.

The term *stock enhancement* is originally derived from efforts to augment wild fish sub-populations, or "stocks," by releasing cultured fishes into aquatic environments. Stocking cultured organisms is one of the tools available for managing aquatic natural resources. It has been used with varying degrees of success to help increase abundance of habitat- or recruitment-limited stocks to help restore depleted populations, augment fisheries and help recover threatened or endangered species. There has been much debate over the effectiveness of stock enhancement as a fisheries management tool. However, most of the scientific evaluation of stocking is quite recent [3], as is a code of responsible practices that help guide effective application [4–6], and marine fisheries enhancement is finally poised for effective use.

In the USA, from the 1880s through the early 1950s, stocking hatchery-reared marine fishes was a principal approach used by the US Fish Commission (renamed Bureau of Fisheries in 1903, Bureau of Commercial fisheries in 1956, and later the National Marine Fisheries Service) for maintaining fishery stocks. But by the 1950s the practice of stocking marine fishes to manage US fisheries was curtailed for lack of evidence of its effectiveness in fisheries management [7]. Stocking was replaced by harvest management to control total catch and sustain fisheries. Stocking of freshwater habitats continued (particularly with salmonids into rivers), although the scientific basis for many of the management decisions needed for stocking salmonids was clearly lacking and did not begin to be addressed until the mid-1970s.

In the decade following 1975, scientists began to evaluate survival and fishery contributions of stocked salmon enabled by advances in fish tagging technology [8, 9]. Quantitative evaluation of marine fish stocking began in earnest in the 1980s and 1990s. The science underlying fisheries enhancement has since evolved to the point where, in some situations, stocking can be a useful fishery management tool to help restore depleted stocks and increase abundance in recruitment-limited fisheries [6]. Effective use of enhancement, though, requires full integration with harvest and habitat management, and a good understanding by stakeholders and resource managers of the opportunities where enhancement can be used successfully as well as its limitations [5, 6]. Principles for guiding the successful use of marine fisheries enhancement to help sustain aquatic resources are now being employed to design new enhancements and reform existing efforts. What follows is a brief overview of those principles and progress made in using hatchery-reared organisms to help sustain marine resources.

Introduction

Marine fisheries enhancement is happening around the world and in some countries on a massive scale (e.g., China). However, in many countries the careful assessment of genetic and ecological risks is lagging behind implementation, putting wild stocks, the seafood supply, and sport fisheries at risk. The science of marine enhancement is still in its infancy compared to other fields of fisheries science, but now shows good potential to (1) increase fishery yield beyond that achievable by exploitation of the wild stock alone, (2) help restore depleted stocks, (3) provide protection for endangered species, and (4) provide critical information on the natural ecology, life history and environmental requirements of valuable marine species.

Stock enhancement has often been used as a generic term referring to all forms of hatchery-based fisheries enhancement. Bell et al. [3] and Lorenzen et al. [6] classified the intent of stocking cultured organisms in aquatic ecosystems into various basic objectives. Together, they considered five basic types, listed here from the most production-oriented to the most conservation-oriented:

1. Sea ranching – *recurring release of cultured juveniles into unenclosed marine and estuarine environments for harvest at a larger size in "put, grow, and take" operations*. The intent here is to maximize production for commercial or recreational fisheries. Note that the released animals are not expected to contribute to spawning biomass, although this can occur when harvest size exceeds size at first maturity or when not all the released animals are harvested.
2. Stock enhancement – *recurring release of cultured juveniles into wild population(s) to augment the natural supply of juveniles and optimize harvests by overcoming recruitment limitation in the face of intensive exploitation and/or habitat degradation*. Stock enhancements can increase abundance and fisheries yield, supporting greater total catch than could be sustained by the wild stock alone [10]. However, such increases may be offset, at least in part, by negative ecological, genetic, or harvesting impacts on the wild stock component. Stock enhancements tend to attract greater numbers of fishers, which can offset expected increase in each individual's catch-per-unit-effort (CPUE) [5, 11].
3. Restocking – *time-limited release of cultured juveniles into wild population(s) to restore severely depleted spawning biomass to a level where it can once again provide regular, substantial yields* [12]. Restocking requires release number to be substantial relative to the abundance of the remaining wild stock, and close ecological and genetic integration of wild and cultured stocks, combined with very restricted harvesting [6].
4. Supplementation – *moderate releases of cultured fish into very small and declining populations, with the aim of reducing extinction risk and conserving genetic diversity* [13, 14]. Supplementation serves primarily conservation aims and specifically addresses sustainability issues and genetic threats in small and declining populations [6].
5. Reintroduction – *involves temporary releases with the aim of reestablishing a locally extinct population* [15]. Continued releases should not occur, as they could interfere with natural selection in the newly established population. Fishing should also be restricted to allow the population to increase in abundance rapidly [6].

Scientific development of marine fisheries enhancement was lacking throughout most of the twentieth century. Although stocking cultured marine fishes began in the nineteenth century, the technology was limited to stocking only eggs and larvae. There were no published accounts of the fate of released fish until empirical studies of anadromous salmonids began to be published in the mid-1970s [16, 17], followed by the first studies (published in English) of stocked marine invertebrates in 1983 [18, 19] and marine fishes in 1989 [20].

During the past two decades, the field of marine fisheries enhancement has advanced considerably. Science in this field is rapidly growing, in part because of critical examination and debate about the efficacy of enhancement and the need for quantitative evaluation (e.g., [21, 22]), and in part because of advances made in aquaculture, genetics, tagging, and fishery modeling technologies, which have enabled quantitative studies and predictions of stocking effects. A clear process has emerged for developing, evaluating, and using enhancement [4–6]. Together, this process and the rapid growth of knowledge about enhancement effects should enable responsible and effective use of enhancement in marine fisheries management and ocean conservation.

Scientific Development of Marine Fisheries Enhancement

Scientific and Strategic Development

Since 1989, progress in marine fisheries enhancement has occurred at two levels – scientific advances and adoption of a careful and responsible approach to planning and organizing enhancement programs and manipulating abundance of marine species using aquacultured stocks. Much of the progress made in the 1990s was scientific and involved an expansion of field studies to evaluate survival of released fish and improve the effectiveness of release strategies. The earliest studies on effectiveness of stocking *marine* fishes, published in English in the scientific literature, were in Japan [20, 23–26] and Norway [27–31], followed by studies in the USA [32–39], and Australia [40]. Progress made with invertebrates is well covered by Bell et al. [12].

Following the initial publications of scientific studies of marine fish enhancement, the number of peer-reviewed publications and symposia in this field began to escalate ([41–52], and see abstracts in [53]). It is now clear that stocking marine organisms can be an effective addition to fishery management strategies, but only when certain conditions are met. For stocking to be productive and economical, and help ensure sustainability of wild stocks, careful attention must be given to several key factors and stocking must be thoroughly integrated with fisheries management [6]. It is clear that stocking can be harmful to wild stocks if not used carefully and responsibly.

Aside from scientific gains in this field, the other level of progress made in the past two decades has been the evolution of a strategic "blueprint" for enhancements, such as the principles discussed in "a responsible approach to marine stock enhancement" [4, 6]. By the early 1990s, salmon enhancement in the US Pacific Northwest, which had been underway for a century, was beginning to incorporate reforms that were needed to improve efficiencies and protect wild stocks from genetic hazards that can lead to loss of genetic diversity and fitness. Concerns had been mounting over uncertainty about the actual effectiveness of salmon hatcheries and impacts on wild stocks. Concerns about wild stock impacts were twofold, including ecological effects of hatchery fish, such as competitive displacement, and genetic issues, such as translocation of salmon stocks, domestication and inbreeding in the hatchery and associated outbreeding depression, and loss of genetic diversity related to hatchery breeding practices (e.g., [54, 55]). Meanwhile, special sessions on marine stock enhancement began appearing at major fisheries and mariculture conferences in the early 1990s [41–44]. These sessions took a sharp turn from past approaches, where the principal focus in conference presentations about stock enhancement had been mainly on Mariculture research topics alone. The conveners of the special sessions on stock enhancement in the 1990s recruited presenters who worked on evaluating the effects and effectiveness of stocking hatchery organisms into the sea and interactions of hatchery and wild stocks. The special sessions focused on the "questions of the day" in marine enhancement and fostered debate in the marine enhancement research community about many of the reform issues being considered in salmon enhancement. The early 1990s was a period of rapid developments in enhancements, characterized by engagement of multiple scientific disciplines in a field that had previously been guided largely by a single discipline – aquaculture.

The salmon experience and reforms underway in salmon enhancement made it clear that a careful and multidisciplinary approach was needed in the development and use of marine enhancement. Many involved in developing new marine fisheries enhancement projects were paying close attention to the debate that had emerged over salmon hatcheries. Following the 1993 special session on "fisheries and aquaculture interactions" held at a mariculture conference in Torremolinos, Spain [44], several of the presenters (including scientists from Japan, Norway, the USA, and Italy [United Nations Food and Agriculture Organization, FAO]) met and formed an "International Working Group on Stock Enhancement," and affiliated the workgroup with the World Aquaculture Society. At that inaugural working group meeting, a decision was made to publish a platform paper to frame the question, "what is a responsible approach to marine stock enhancement?" This paper was presented at the 1994 American Fisheries Society symposium, "Uses and Effects of Cultured Fishes in Aquatic Ecosystems," and published in the 1995 peer-reviewed symposium proceedings [4]. The paper recommended ten principles for developing, evaluating, and managing marine stock enhancement programs. The Responsible Approach paper afforded a model for developing and managing new enhancement programs and refining existing ones. It has also helped frame research questions in the emerging science of marine fisheries enhancement.

The International Working Group on Stock Enhancement (IWGSE) was instrumental in advancing the science of marine fisheries enhancement in the 1990s. The working group focused primarily on highlighting ongoing stock enhancement research around the world and fostering awareness of the Responsible Approach in their publications and presentations. International awareness and new research in the field was aided by the broad international makeup of the working group. Membership grew and soon included scientists from Australia, Canada, China, Denmark, Ecuador, Italy, Japan, Norway, Philippines, Solomon Islands, Spain, the UK, and the USA. Initially, the primary communication vehicle used by the working group was the special sessions on stock enhancement, which it planned and convened annually in various countries at the international conference of the World Aquaculture Society. The working group promoted a synergy among its members and the influence of the group expanded as members planned additional workshops and symposiums in their own countries and brought IWGSE scientists into the planning process.

The period 1990–1997 was a fertile time that gave birth to a rapid expansion of science in marine fisheries enhancement, which continues to this day, aided since 1997 in large part by the International Symposium on Stock Enhancement and Sea Ranching (ISSESR). The first ISSESR, held in 1997 in Bergen, Norway, was the brainchild of the Norwegian PUSH program (Program for Development and Encouragement of Sea Ranching) and the Norwegian Institute of Marine Research (IMR). In 1995, IMR scientists invited IWGSE scientists to become involved in the International Scientific Committee charged with planning the program for the first ISSESR. The first ISSESR, and the series of follow-up symposia that it launched (see www.SeaRanching.org), have encouraged and brought about fundamental advancements in the field of marine enhancement – by networking the scientists working in this specialized field, highlighting their work at the ISSESR, and publishing their peer-reviewed articles in the symposium proceedings. The 3–5 day ISSESR has now become a regular scientific symposium event, hosted by a different country every 4–5 years. Following the first ISSESR in Bergen [47], subsequent symposiums in the series were held in Kobe, Japan in 2002 [49], in Seattle, USA in 2006 [52], and in Shanghai, China in 2011 [53]. The fifth ISSESR will be held in Sydney, Australia in 2015 or 2016. Inquiries from scientists in different countries interested in hosting the sixth one are already being received by the organizing group. Following the first ISSESR, the IWGSE scientists continued the efforts they started in the working group through their involvement in the International Scientific Committees for the ISSESR and steering committees for other stock enhancement symposia (e.g., [46, 48, 51]). In 2010, a refined and updated version of the Responsible Approach was published [6] and presented at the fourth ISSESR.

As in any new science, lack of a paradigm and consensus on the key issues retard progress. The ISSESR and other marine enhancement symposia and working groups have helped to place scientific focus on critical uncertainties and communicate results of new science in this field at symposiums and in the scientific literature. They have also provided a forum for debate on the issues, and increased networking of scientists, resource managers, students, and educators working in this field worldwide. The focus on key issues is nurturing this new field of science.

Technological and Tactical Constraints

Although marine enhancements do show promise as an important tool in fisheries management, why has this field taken so long to develop and why have marine enhancement programs often failed to achieve their objectives? The scientific development of marine fisheries enhancement has long been impeded by lack of the technologies needed to evaluate effects of stocking cultured fish. Although marine enhancements began in the 1880s, until the advent of the coded-wire tag in the mid-1960s [8], there was no way to identify treatment groups and replicates in experimental releases of juvenile cultured fish [56]; and quantitative marking methods for multiple experimental groups of postlarvae and very small juveniles (<50 mm in length) came much later (e.g., [57]). To make matters worse, scientific development of marine enhancement was also stymied by lack of adequate technology for culturing marine fishes. Rearing methods for larval and juvenile marine fishes, many of which require live feeds during the larval stage, remained undeveloped until the mid- to late 1970s, when breakthroughs finally began to be achieved in rearing a few marine species past metamorphosis [58]. By the mid-1980s mass production of juveniles had been achieved for several species of marine fishes. Even today, though, many marine fishes cannot yet be cultivated to the juvenile stage in the quantities needed for stocking. Without the availability of juveniles grown to a wide range of sizes, fundamental questions about density dependence, hatchery-wild fish interactions and cost-yield efficiency of size-at-release and other release variables cannot be addressed in field experiments. Thus, even the basic technologies needed to develop and understand the potential of marine enhancement have been unavailable until relatively recent times for some fishes and have yet to be developed for others.

Technology has not been the only constraint to successful development of marine fisheries enhancement. The effective use of stocking cultured marine organisms in fisheries management has been hindered by lack of understanding of the effect of releases on fish population dynamics and a lack of related, quantitative assessment tools [10]. Moreover, there has been a lack of essential governance and fisheries management considerations in planning, designing, implementing, and evaluating enhancement programs [6, 59]. A symptom of this is the relentless concern among stakeholders and hatchery managers alike about the numerical magnitude of fish released, rather than on the effective contribution of the hatchery program to fisheries management goals. Certainly, a hatchery needs to meet some release quotas, but the numbers of fish released is a misleading statistic for gauging success or comparing effectiveness among enhancement programs. Yet, from the very beginning, progress has been judged by the number of eggs, yolk-sac larvae or juveniles stocked, rather than by the number of fish added to the catch or to spawning stock biomass. The thinking behind this approach apparently is "grow and release lots of hatchery fish and of course they'll survive and add to the catch," without realizing the need to optimize release strategies (e.g., [39, 60, 61]) (e.g., to know what size-at-release, release habitat and release magnitude combination has the greatest impact on population size, fishery yields, and economics), or that the impact from stocking could in fact be a negative one on wild stocks (such as replacement of wild fish by hatchery fish) if certain precautions are not taken. This attitude has been pervasive and exists even today among many stakeholders and enhancement administrators. In fact, research now shows that survival and recruitment to the fishery following hatchery releases is a complex issue that requires much greater understanding about the fishery, hatchery fish performance, and biological and ecological factors in the wild than simply "the catch is down, thus releasing large numbers of fish will bring it back up." And quite often large release magnitudes are achieved by releasing millions of postlarvae, rather than fewer but larger juveniles. But releases of postlarvae alone may be effective, yet can also be totally ineffective, depending on conditions at the release site [62].

The key to successful use of stocking is to plan enhancement programs from a fisheries/resource management perspective, using a broad framework and scientific approach [6, 59]. The probability of achieving effective results is greatly increased when stakeholders are engaged from the outset in planning *new* programs, using a framework that is structured, multilayered, participatory, and makes good use of science, to design, implement, and analyze enhancement fisheries systems [6]. Incorporating the key principles in the Responsible Approach into the frameworks of *existing* programs as well is likely to improve performance.

Responsible Approach to Marine Fishery Enhancement

In retrospect, the slow development of marine fish culture (a century behind salmonid aquaculture) has helped marine stock enhancement programs avoid some of the mistakes of the past made with salmon stock enhancement, where lack of understanding of genetic issues during most of the twentieth century led to inadvertent domestication and inbreeding in salmon hatchery populations, leading to reduced fitness in wild stocks. Marine finfish juvenile production technology lagged behind freshwater and anadromous fish culture by a century. Thus, mass release into the sea of juvenile marine fishes large enough to survive and enter the breeding population did not begin until the 1980s. The relatively recent capabilities to conduct marine fisheries enhancement emerged at about the same time that geneticists realized that hatchery practices with salmonids (1) could reduce genetic diversity in the hatchery and ultimately, enhanced wild stocks, owing to inadequate broodstock management, (2) have caused translocations of salmon genes into environments where they are less fit, and (3) have contributed to loss of local adaptations in the wild population. Today, population genetics is much better understood and broodstock genetics and hatchery practices can be better managed to address these concerns (e.g., [63–65]). Thus, marine enhancement programs need careful guidance from qualified geneticists. The Puget Sound and Coastal Washington Hatchery Reform Project in the USA has been instrumental in reforming salmon enhancements [66]. This group affords a model for managing enhancement hatcheries in the twenty-first century.

As progress was being made in the early 1990s to better understand the genetic structure of stocks and how to manage genetics in hatcheries, realizing the need for reform in approaches to enhancing non-salmonids was just beginning. In the mid-1990s, Cowx [67], for enhancements in freshwater systems, and Blankenship and Leber [4], for enhancements in marine and estuarine systems, published papers calling for a broader, more systematic, reliable, and accountable approach to planning stock enhancement programs. Prompted both by the salmonid hatchery reform movement and by the WAS IWGSE, the ten principles presented in Blankenship and Leber ([4] Table 1) gained widespread acceptance as the "Responsible Approach" to stocking marine organisms and provided a platform for subsequent discussions on planning, conducting, and evaluating marine enhancements (e.g., [6, 12, 22, 51, 52, 68–70]). Since 1995, the awareness of the Responsible Approach has steadily increased and has helped guide hatchery and reform processes for marine enhancements worldwide [11, 36, 37, 39, 60, 62, 69–90].

Marine Fisheries Enhancement, Coming of Age in the New Millennium. Table 1 The ten principles of a responsible approach to marine stock enhancement [4]

1	Prioritize and select target species for enhancement by ranking and applying criteria for species selection
2	Develop a management plan that identifies how stock enhancement fits with the regional plan for managing stocks
3	Define quantitative measures of success to track progress over time
4	Use genetic resource management to avoid deleterious genetic effects on wild stocks
5	Implement a disease and health management plan
6	Consider ecological, biological, and life history patterns in forming enhancement objectives and tactics; seek to understand behavioral, biological, and ecological requirements of released and wild fish
7	Identify released hatchery fish and assess stocking effects on the fishery and on wild stock abundance
8	Use an empirical process for defining optimal release strategies
9	Identify economic objectives and policy guidelines, and educate stakeholders about the need for a responsible approach and the time frame required to develop a successful enhancement program
10	Use adaptive management to refine production and stocking plans and to control the effectiveness of stocking

The Responsible Approach provides a conceptual framework and logical strategy for using aquaculture technology to help conserve and increase natural resources. The approach prescribes several key components as integral parts of developing, evaluating and managing marine fisheries enhancement programs. Each principle is considered essential to manage enhancements in a sustainable fashion and optimize the results obtained [4, 6].

A major development since the publication of the original "Responsible Approach" has been increasing interest from fisheries ecologists in understanding and quantifying the effects of hatchery releases from a fisheries management perspective. This has led to the development of fisheries assessment models that can be used to evaluate stocking as a management option alongside fishing regulations [5, 10]. At the same time, approaches to fisheries governance underwent major changes that allow enhancements to become more integrated into the management framework and in some cases, were driven by interest in enhancement approaches [59].

Walters and Martell [5] discuss four main ways that a marine enhancement program can end up causing more harm than good: (1) the replacement of wild with hatchery recruits, with no net increase in the total stock available for harvest (competition/predation effects); (2) unregulated fishing-effort responses to the presence of hatchery fish that cause overfishing of the wild stock; (3) "overexploitation" of the forage resource base for the stocked species, with attendant ecosystem-scale impacts; and (4) genetic impacts on the long-term viability of the wild stock. They stress that it is critical to monitor the impacts of enhancement as the program develops to have evidence in hand if debate about the efficacy of the program does surface. To help guide developing programs, they provide and discuss a "Code of Responsible Conduct" as critical steps in marine fisheries enhancement program design (Table 2).

Marine Fisheries Enhancement, Coming of Age in the New Millennium. Table 2 Code of responsible conduct for marine stock enhancement [5]

• Make certain that management priorities and acceptable trade-offs are absolutely clear
• Do careful stock assessments to show that the target stock is recruitment overfished or can no longer rear successfully in the wild
• Show that enhanced fish can recruit successfully in the wild
• Show that total abundance is at least initially increased by the hatchery fish contribution
• Show that fishery regulations are adequate to prevent continued overfishing of the wild population, unless there has been an explicit decision to "write off" the wild population
• Show that the hatchery production system is actually sustainable over the long run, when it is to be a permanent component of the production system

In 2010, Lorenzen, Leber, and Blankenship [6] published an updated version of the Responsible Approach to refine the original key principles and include five additional ones (Table 3). The key principles added in the updated version bring stakeholders more firmly into the planning process; place much stronger emphasis on a-priori evaluation of the potential impact of enhancements using quantitative models; place marine fishery enhancements more firmly within

Marine Fisheries Enhancement, Coming of Age in the New Millennium. Table 3 The updated responsible approach (From [6])

Stage I: Initial appraisal and goal setting	
1	Understand the role of enhancement within the fishery system [*new*]
2	Engage stakeholders and develop a rigorous and accountable decision making process [*new*]
3	Quantitatively assess contributions of enhancement to fisheries management goals
4	Prioritize and select target species and stocks for enhancement
5	Assess economic and social benefits and costs of enhancement
Stage II: Research and technology development including pilot studies	
6	Define enhancement system designs suitable for the fishery and management objectives [*new*]
7	Design appropriate aquaculture systems and rearing practices [*new*]
8	Use genetic resource management to maximize effectiveness of enhancement and avoid deleterious effects on wild populations.
9	Use disease and health management
10	Ensure that released hatchery fish can be identified
11	Use an empirical process for defining optimal release strategies
Stage III: Operational implementation and adaptive management	
12	Devise effective governance arrangements [*new*]
13	Define a management plan with clear goals, measures of success, and decision rules
14	Assess and manage ecological impacts
15	Use adaptive management

the context of fishery management systems; emphasize design of appropriate aquaculture rearing systems and practices; and incorporate institutional arrangements for managing enhancements. Lorenzen et al. [6] provide comprehensive discussions for each of the 15 key principles listed in Table 3. Readers are urged to consult Lorenzen et al. [6] for additional detail, as it is beyond the scope, here, to repeat their discussions of each principle.

The 15 principles in the updated Responsible Approach include the broad range of issues that need to be addressed if enhancements are to be developed or reformed responsibly [6]. Clearly, marine enhancement programs are multidisciplinary and their effective use requires specialist knowledge and skills from diverse fields (Table 4). Forming interdisciplinary teams of the various specialists required is an important factor in employing the Responsible Approach in developing, reforming, and executing marine enhancements. For effective design of enhancement programs, specialists in each area of expertise listed in Table 4 should be included in the planning teams.

It should be clear that without a careful monitoring system in place, marine enhancements simply cannot be managed. Monitoring is essential to understand the impacts of enhancement, to manage release strategies so that they are efficient and designed well enough to achieve the goals of the program, to protect against misuse of stocking (as discussed in 5 and 6), resulting in harm to wild stocks, and to document success or failure in meeting enhancement program objectives. Walters and Martel [5] list several key monitoring requirements for managing fishery enhancements well: (1) mark all (or at least a high and known proportion of) fish released from hatcheries; (2) mark as many wild juveniles as possible at the same sizes/locations as hatchery fish are being released; (3) experimentally vary hatchery releases over a wide range from year to year and from area to area, probably in on/off alternation (temporal blocking) so as to break up the confounding of competition/predation effects with shared environmental effects; (4) monitor changes in total recruitment to, production of, and fishing effort in impacted fisheries, not just the percentage contribution of hatchery fish to production; (5) monitor changes in the fishing mortality rates of both wild and hatchery fish directly, through carefully conducted tagging programs that measure short-term probabilities of capture; and (6) monitor reproductive performance of hatchery-origin fish and hatchery-wild hybrid crosses in the wild. Sound management-action design and monitoring is the essence of adaptive management [91] and adaptive management enables refinements, progress, and success in marine enhancement programs [4, 6, 11, 92].

Marine Fisheries Enhancement, Coming of Age in the New Millennium. Table 4 Key areas of expertise needed in marine fisheries enhancement

• Fisheries science
• Fisheries management
• Adaptive management
• Marine aquaculture
• Population genetics
• Aquatic animal health
• Population ecology
• Behavioral ecology
• Community ecology
• Resource economics
• Social science and institutional analysis and design
• Statistics and experimental design
• Tagging technology
• Communications and outreach

Marine fisheries enhancement is a powerful tool that requires careful and interdisciplinary planning to control its effects. The process of transforming marine enhancement from an idea before its time into an effective resource management and sea ranching tool involves adopting a clear prescription for responsible use. As marine enhancement comes of age in this new millennium, agencies and stakeholders have a growing library of protocols for enhancement at their disposal and the responsibility to use them. The Responsible Approach and Code of Responsible Conduct provide healthy prescriptions for controlling the outcome of enhancements. These principles need to be adopted and used well, in order to increase and ensure the readiness of this tool to aid in conservation and to increase fishery yields when it is needed. Growth in human population size is fast approaching a critical level, and much greater attention will be placed in this century on obtaining food from the sea [1]. It is unwise to not be ready with marine enhancement to help sustain depleted, threatened, and endangered species, help maintain wild stocks in the face of increasing fishing pressure, help sustain sports fisheries, and help increase fishery yields.

Legacy from the Past

Allure of a Quick Fix

Marine enhancement programs are often seen as a "quick fix" for a wide variety of problems in marine resource management. At best, they may be an important new component of marine ecosystem management; if not implemented responsibly, though, they may lull fishery managers into false confidence and thus lead to inaction and delay in the development of other fisheries management and restoration programs [5, 6].

Although marine fisheries enhancement is certainly not a quick fix, it can be a powerful tool for resource management when conditions warrant the use of this tool and if the time and care needed are taken to develop enhancement programs well. Unfortunately, the allure of a quick fix has often prompted stakeholders and managers to skip or ignore several elements needed to allow those programs to succeed, leading to wholesale failure of such efforts. The field of marine fisheries enhancement is littered with examples of enhancement projects that failed to achieve their potential for lack of a careful enough or quantitative approach (e.g., see accounts discussed in [7, 21, 62, 72, 93–95]). Most of the failures can be traced back to attempts to use enhancements when they were not warranted or failure to consider several, if not most, of the principles now incorporated in the "Responsible Approach" and "Code of Responsible Conduct" for marine fisheries enhancement.

Isolation from the Fisheries Science Community

Historically, marine fisheries enhancements have been conducted more or less isolated from other forms of fisheries management. Enhancement hatcheries have often been promoted by stakeholders and government mandates without the necessary funding or authorization behind them to do much more than produce and release fish without funds for monitoring impacts and adaptive management needed to increase the effectiveness of enhancements. Such programs are often built and implemented from a vantage point within resource management agencies that has little or no connectivity with the existing fishery management process. This has stymied development of this field in two ways – first, by compelling hatcheries to operate within resource management agencies largely independent from stock assessment and fisheries monitoring programs, or even worse, within different agencies altogether. Second, such isolation has fostered development of a production-oriented operational mode, and thwarted development of an enhancement-oriented mode [92].

Part of this isolation from fishery management also stems from the poor track record of the early marine hatcheries as an effective way to recover depleted fish stocks, coupled with the lack of scientific development of marine fisheries enhancement for so long into the twentieth century. This has understandably led to bias against fishery enhancements. Many of today's fishery scientists have been schooled to understand that stock enhancement has not worked, based in part on the lingering legacy from past failures and in part on lack of awareness of new marine fisheries enhancement science, as few citations have yet appeared in fisheries science textbooks. With many of the scientific achievements in fisheries enhancement having occurred only over the past decade or so, this is understandable. But in light of the need to couple fisheries enhancement with fisheries management systems, lack of awareness of progress in this field is an obstacle that may be resolved only by compilation of more and more success stories over time. Thus, it is imperative that existing and developing enhancement programs alike incorporate modern concepts about how to plan and conduct enhancements so they are enabled for success.

Progress in Marine Fisheries Enhancement

Lessons Learned from Marine Enhancement Programs

Much progress has now been made in understanding how to manage enhancement more effectively.

Bartley and Bell [96] considered progress made from three decades of stocking initiatives and summarized and discussed lessons learned. These are listed here, below [96], with a brief clarification or caveat on each.

Deciding When and How to Apply the Release of Cultured Juveniles

1. Objective assessment of the need for releases is crucial –and requires an evaluation of the status of the fishery, modeling of stocking impact to determine if stocking can help achieve the goals, coupled with consideration of whether there are recruitment limitations and adequate habitat available for stocking.
2. Releases of cultured juveniles for restocking and stock enhancement need to be made at the scale of self-replenishing populations –releases will not be effective unless the spatial extent of target populations has been identified; thus prior to conducting releases of hatchery organisms, clear identification of genetically discrete stocks should be determined.
3. There are no generic methods for restocking and stock enhancement –largely because of wide variation in life history among different species and variation in ecological conditions among release sites.
4. Very large numbers of juveniles are often needed for effective stock enhancement –this is particularly so for offshore stocks, which can be comprised of a huge number of individuals; more modest releases may suffice for localized enhancement of inshore stocks or those comprised of multiple stocks that occur on relatively small scales.
5. Large areas are needed for stock enhancement of some species –and this can result in user conflict, particularly for sea ranching, where large areas are leased and protected by the enhancement program (e.g., [97]); in other cases, limited dispersal of adults and larvae indicates stocking in smaller areas can be effective, for example, common snook along Florida's Gulf Coast [98].
6. Invertebrates offer good opportunities for restocking and stock enhancement –because invertebrates are often comprised of self-recruiting populations that occur at small scales.

Integrating Interventions with Other Management Measures

7. Problems that caused lower production must be addressed before release of juveniles – particularly in the case of degraded, lost, or insufficient habitat. With better management of the wild resources, the scope for augmentation of total production declines; enhancement becomes a very site specific tool when habitat has been lost, or something needs rebuilding, or there are species of particularly high value [94].
8. Biotechnical research must be integrated with institutional and socio-economic issues – ownership rights and control and use of enhanced stocks need to be well understood by the greater institutional, social, economic, and political environment [99].
9. Successful stock enhancement programs are often run by cooperatives and the private sector – where there is increased incentive in sharing the costs of fisheries enhancement.
10. The costs and time frames involved in restocking programs can be prohibitive – hatchery costs, which can be considerable, are particularly difficult to bear in smaller countries and developing countries.

Monitoring and Evaluation

11. Development of cost-effective tagging methods is critical to efficient evaluation of stock enhancement – refining and monitoring the effects and effectiveness of marine enhancements cannot be done without a way to distinguish hatchery from wild stocks and distinct release groups.
12. Large-scale releases of hatchery-reared juveniles can affect genetic [fitness] of wild populations – genetic hazards can be caused by hatchery-wild fish interactions and these need to be minimized.

Reducing the Cost of Juveniles

13. Costs of stocking programs can be reduced by "piggybacking" production of juveniles for release on existing aquaculture – this could reduce or eliminate the need for expensive new hatchery construction for enhancement programs, as long as appropriate broodstock management protocols are in place for conserving wild-stock genetics.
14. Wild [postlarvae] can provide an abundant, low-cost source of juveniles for stock enhancement

programs – this can sometimes be an effective way to reduce costs and eliminate genetic issues; successful scallop enhancement in Japan is based on collection of wild seed stock.

15. The costs of restocking can be reduced greatly for some species by relocating adults to form a viable spawning biomass – rebuilding spawning aggregations by concentrating broodstock can be effective for depleted stocks with limited larval dispersal, but care must be taken to avoid comingling different stocks (i.e., avoid translocation of exogenous genes).

Improving Survival in the Wild

16. Predation is the greatest hurdle to survival of released juveniles – care must be taken to understand ecology of the species and ecosystem at the release site and pilot experiments are needed to develop optimal release strategies to maximize survival.
17. Excessive releases of juveniles cause density-dependent mortality – density has a strong effect on growth and survival in the wild; planning release magnitude must take into account the carrying capacity at release locations. This requires adaptive management and an experimental framework for releases.
18. Small-scale experiments to test methods for releasing juveniles can give misleading results – "commercial scale" releases are needed to test assumptions made from small-scale release experiments.
19. Good survival of released juveniles at one site is no guarantee that the methods can be transferred to other sites – stocking effectiveness will vary with release location and what works at one site may not be effective at another.

Other Manipulations to Increase Abundances

20. Artificial habitats can be used to increase the carrying capacity for target species – and may enable increased production at release sites where there are resource (food, refuge, space) limitations.
21. Yields of some species can be increased by providing suitable settlement habitat and redistributing juveniles from areas of heavy settlement – for example, redistribution can be used to reduce density effects and increase probability of successful recruitment when moved to a location with greater availability of food, refuge, or settlement habitats. But care must be taken to avoid genetic hazards associated with comingling stocks.

Examples of Progress Made in Marine Enhancement

As science and constructive debate have advanced in this field, there are many signs of progress. Some explicit examples of progress made in marine enhancement over the past couple of decades are presented below, ranging in scale from local experimental investigations of release strategies and density-dependent effects on hatchery and wild stocks (e.g., [100]) to documented replenishment impact in large-scale enhancement efforts (e.g., [101, 102]). This is but a sample of examples and is by no means a comprehensive list. There are many more examples in the peer-reviewed proceedings from the ISSESR and other stock enhancement conferences [41–53] and other journal articles.

1. Adoption of a science-based responsible approach to marine stock enhancement has now become widespread, resulting in a much more assessment-driven and precautionary approach than ever before (a few examples include Refs. [4, 6, 10, 12, 20, 22, 27–29, 33, 37–39, 59–61, 68, 69, 72, 75, 84, 86, 87, 89, 96, 103–106]). This has been enabled, in part, by advances in tagging technology (e.g., [8] and see examples in [9, 56]) and in development of new marine aquaculture technologies that can now provide juvenile fishes for marine enhancement research.
2. Networking of Scientists involved in this rapidly advancing field has been fostered by various symposia and working groups, for example, the World Aquaculture Society Working Group on Stock Enhancement and the scientific committees for the International Symposium on Stock Enhancement and Sea Ranching (www.SeaRanching.org).
3. There is a much better appreciation of the importance of managing marine fishery enhancements from a fisheries management perspective (e.g., [6, 59, 107]).
4. New tools are available for modeling stock enhancement effects and effectiveness [10, 82, 108–110].

5. At least two experimental field studies have now been conducted to evaluate density-dependent interactions of stocked hatchery and wild fish; these provide evidence that increased production can be achieved in juvenile nursery habitats without displacing wild fish, but not necessarily without displacing some of the hatchery fish [33, 100].
6. There is now clear evidence and a prescription of techniques for improving post-release survival (often with a doubling effect or more) of stocked marine fishes, and optimizing release strategies to maximize stocking efficiency and control impacts (e.g., [26, 36, 37, 39, 60–62, 70, 72, 100–115]). There is also ample evidence that in habitats with limited carrying capacity or intense predation, regardless of release strategy used, little can be done to improve survival of hatchery fish and stocking simply cannot increase production [106, 116, 117].
7. It is now fairly clear that marine enhancements may be cost effective only if (a) the supply of recruits is generally limiting, (b) there is adequate habitat to support an increased supply of juveniles, (c) cultured juveniles represent a large portion of recruitment, (d) fishing is regulated appropriately, and (e) other management measures (catch regulations and habitat restoration) are insufficient to restore catch rates [96].
8. Stock enhancement of some species of marine finfish has been successful at the scale of large bays, for example, Hirame flounder and red sea bream in Japan [72, 106] when there is sufficient carrying capacity at release sites. Carrying capacity varies considerably among release sites, and thus must be evaluated and taken into account using monitoring and adaptive management for each release site.
9. Scallop sea ranching has been a large success in Japan, New Zealand, and China, where property rights and large ocean leases have created strong incentives for careful management by fishermen and owners of the sea ranching operations [72, 101, 102, 118]. For example, near Dalian, China, Zhangzidao Fishery Group leases 2,000 km^2 of ocean-bottom-to-ocean-surface for sea ranching. In 2010, Zhangzidao harvested an average of 150 t/day of ocean scallops from their sea ranching operations (over 50,000 t/year) (Wang Qing-yin, personal communication 2011).
10. Property rights have also provided incentives for bivalve culture in the State of Washington, USA, where clam sea ranching operations have remained economically and environmentally sustainable for over three decades [119].
11. Pilot experiments with black bream in an Australian estuary have documented quite good survival and recruitment to the fishery. The latest phase of this project reveals strong rationale for long-term monitoring of enhancement impact [87, 120].
12. Restocking success with red drum in a South Carolina estuary [77, 121]. Pilot experiments revealed surplus productive capacity in the Ashley River in South Carolina, where fishery landings of red drum were doubled over a few years.
13. Pilot experiments to evaluate blue crab enhancement potential in Maryland and Virginia led to improvements in traditional fishery management, with information learned through stocking research [70, 114]. Pilot experiments can be used to provide critical information on the natural ecology, life history, and environmental requirements of valuable marine species [122].
14. Perhaps the largest scale enhancement success for fishes is Japanese chum salmon restocking – a special tool for a circumstance in which the habitat had almost totally been lost [94].

Future Directions

Over the past two decades, there has been a rapid expansion of knowledge about marine fisheries enhancement systems and the effects and effectiveness of stocking a wide variety of marine organisms for sea ranching, stock enhancement and restocking. Many gaps in knowledge have now been filled. Well thought out approaches now provide a roadmap for effective use of enhancements. When models show potential for stocking, efforts to deploy marine enhancements can be successful if the principles in the roadmap are carefully employed. The basic reason that marine enhancement programs do not have more of a track record of success stories yet is that implementing them well is a complex endeavor that demands attention to multiple factors spanning many disciplines. Rarely have these been pulled together in an enhancement program. The Hatchery Reform Project in the Pacific Northwest USA, which includes an independent scientific review panel ("Hatchery Scientific Review Group") is a good example [123]. Because of their efforts, salmonid hatchery reforms now underway are bringing many of the principles of the Responsible Approach into play. The Norwegian PUSH program is another good example. In that case, information gained from quantitative assessments of enhancement showed that stocking would not be an economical way to enhance cod in Norway, thus saving years of wasteful spending that could have occurred there, had monitoring and adaptive management not been a central part of the enhancement system.

Successful examples of fisheries enhancement are truly group efforts, involving stakeholders, agency officials, and individuals with expertise in the principal sub-disciplines needed. Suffice to say that at this point in time few, if any, marine fisheries enhancement programs have enlisted all of the key elements of the Responsible Approach and Code of Responsible Conduct. But these principles are now well described and laid out in a systematic manner. It is reasonable to expect that if the Responsible Approach is used as the blueprint for planning and executing enhancements, and if the initial appraisal and goal setting stage indicates moving ahead, then there is ample opportunity for success in applying marine fisheries enhancements, as long as dedicated attention is focused on applying each of the key elements.

So how will marine enhancement advance to the next level – emergence of a rapidly growing body of success stories in restocking, stock enhancement, and sea ranching? Listed below are a few factors that are now needed to transition this field to the next level, where marine enhancements are well integrated into resource management systems and used wisely and appropriately.

Enabling Factors for Increasing Successful Marine Enhancements

1. *Greater awareness is needed among all stakeholders of the issues, pitfalls, progress, and opportunities in this field.* The concepts underlying effective enhancements need to be translated into lay language and used to inform stakeholders. This will help all stakeholders recognize the various issues and parameters needed for effective enhancements. Pivotal among stakeholders are public officials who fund enhancement programs, as they need to understand what it takes to develop an effective program or reform existing ones. New enhancement programs that may not be funded well enough to implement all of the key principles in the Responsible Approach would do well to use the results of Stage 1 in Table 3 to document the potential for success, but not proceed beyond Stage 1 until adequate funding is available.
2. *Use of Adaptive management is one of the most important principles for guiding successful enhancement programs.* Active adaptive management [91] is critical for gauging the effectiveness of, improving, and managing fisheries systems in the face of uncertainty. However, it is often dismissed by enhancement programs or given low priority for lack of funding or when enhancement is viewed as a quick fix. But, this important principle is used to optimize release strategies, to identify and deal with ecological or genetic impacts on wild stocks, to refine the enhancement process and identify the results of improvements, to evaluate and improve progress towards goals and objectives, and to monitor and improve economic impact. Active adaptive-management is an essential component of managing enhancement programs; it empowers management teams to understand and control the impacts of enhancements well. Without it, enhancement programs at best rely on hope to achieve their potential (but cannot) and at worst are doomed to failure. Australia is employing active adaptive management principles early in the development stage as part of ongoing work to evaluate enhancement potential for a wide range of species [124].
3. *Adapt the Responsible Approach to local circumstances.* The Responsible Approach is purposely vague on how to implement it. This is partly because not all elements are needed under all situations, but most will be. Fitting the process to particular circumstances is in itself a key part of implementing the Responsible Approach by

engaging the various stakeholders in planning [6]. As progress continues in this field, additional principles will emerge that need to be included, for example, to account for needs of regional fishery management plans in response to climate change.

4. *Seek assistance from established workers in the field.* For new and developing enhancement programs, or existing ones seeking to design and implement reforms, there is a broad and expanding network of workers in this field who could be queried for advice on various enhancement issues. The ISSESR website is a good source for identifying individuals with specific kinds of expertise, by perusing presentation abstracts or locating published proceedings from past ISSESR conferences [125]. If researchers or workers in the field are contacted, but do not have time to provide advice, they usually will help identify others who can.

This entry may help expand awareness among fishery stakeholders, other natural-resource stakeholders, scientists, and fishery managers alike about the pitfalls, challenges, and progress made in using marine hatchery releases as one of the tools in resource management and seafood production. Readers are referred to the articles and symposium proceedings cited herein to gain a better understanding of the issues, lessons learned, and progress.

The debate focused on enhancement is a healthy one, for it is fostering steady improvements and reforms in existing programs, and careful planning and design in new ones. With each advance made, the potential seen by our forefathers to use hatcheries as a tool for recovering depleted stocks, increasing abundance in recruitment-limited stocks, and producing seafood by sea ranching is coming closer to fruition. One of the greatest lessons learned from the past is that the emphasis on expanding hatchery fish production for marine enhancement should not be allowed to take the focus off of the objective – increasing yields in fisheries and recovering stocks in restoration programs. Clearly, marine fisheries enhancement is a strong tool to add to the fishery management toolbox. But only careful analysis of conditions of the wild stock and the fishery will guide when and where it is appropriate to use enhancements in addition to other management options, and when to stop. As Albert Einstein once said, "a perfection of means, and confusion of aims, seems to be our main problem." With the focus shifted to outcomes in marine enhancement programs, the appropriate means should fall into place, aided by healthy debate and prescriptions for a responsible approach to marine fisheries enhancement.

Bibliography

1. Duarte CM, Holmer M, Olsen Y, Soto D, Marbà N, Guiu J, Black K, Karakassis I (2009) Will the oceans help feed humanity? Bioscience 59(11):967–976
2. NOVA (1992) Sex and the single rhinoceros – NOVA examines the high-tech efforts to preserve the world's animal diversity. PBS documentary. NOVA Season 19, Episode 20. Public Broadcasting Service
3. Bell JD, Leber KM, Blankenship HL, Loneragan NR, Masuda R, Vanderhaegen G (eds) (2008) A new era for restocking, stock enhancement and sea ranching of coastal fisheries resources. Special Issue, Rev Fish Sci 16(1–3):402 pp
4. Blankenship HL, Leber KM (1995) A responsible approach to marine stock enhancement. Am Fish Soc Symp 15:167–175
5. Walters CJ, Martell SJD (2004) Fisheries ecology and management. Princeton University Press, Princeton
6. Lorenzen K, Leber KM, Blankenship HL (2010) Responsible approach to marine stock enhancement: an update 2010. Rev Fish Sci 18(2):189–210
7. Richards WJ, Edwards RE (1986) Stocking to restore or enhance marine fisheries. In: Stroud RH (ed) Fish culture in fisheries management. American Fisheries Society, Bethesda, pp 75–80
8. Jefferts KB, Bergman PK, Fiscus HF (1963) A coded-wire identification system for macro-organisms. Nature 198:460–462
9. Blankenship HL, Tipping JM (1993) Evaluation of visible implant and sequentially coded wire tags in sea-run cutthroat trout. North Am J Fish Manag 13:391–394
10. Lorenzen K (2005) Population dynamics and potential of fisheries stock enhancement: practical theory for assessment and policy analysis. Philos Trans R Soc Lond Ser B 260:171–189
11. Leber KM (2004) Marine stock enhancement in the USA: status, trends and needs. In: Leber KM, Kitada S, Blankenship HL, Svåsand T (eds) Stock enhancement and sea ranching: developments, pitfalls and opportunities. Blackwell, Oxford, pp 11–24
12. Bell JD, Rothlisberg PC, Munro JL, Loneragan NR, Nash WJ, Ward RD, Andrew NL (2005) Restocking and stock enhancement of marine invertebrate fisheries. Adv Mar Biol 49:1–370
13. Hedrick PW, Hedgecock D, Hamelberg S, Croci SJ (2000) The impact of supplementation in winter-run Chinook salmon on effective population size. J Hered 91:112–116
14. Hilderbrand RH (2002) Simulating supplementation strategies for restoring and maintaining stream resident cutthroat trout populations. North Am J Fish Manag 22:879–887
15. Reisenbichler RR, Utter FM, Krueger CC (2003) Genetic concepts and uncertainties in restoring fish populations and

species. In: Wissmar RC, Bisson PA (eds) Strategies for restoring river ecosystems: sources of variability and uncertainty in natural and managed systems. American Fisheries Society, Bethesda, pp 149–183

16. Hager RC, Noble RE (1976) Relation of size at release of hatchery-reared coho salmon to age, sex, and size composition of returning adults. Progress Fish Cult 38:144–147
17. Bilton HT, Alderdice DF, Schnute JT (1982) Influence of time and size at release of juvenile Coho Salmon (*Oncorhynchus kisutch*) on returns at maturity. Can J Fish Aquat Sci 39:426–447
18. Appledorn RS, Ballentine DL (1983) Field release of cultured queen conchs in Puerto Rico: implications for stock restoration. Proc Gulf Caribb Fish Inst 35:89–98
19. Appeldorn RS (1985) Growth, mortality and dispersion of juvenile laboratory-reared conchs, *Strombus gigas*, and *S. costatus*, released at an offshore site. Bull Mar Sci 37:785–793
20. Tsukamoto K, Kuwada H, Hirokawa J, Oya M, Sekiya S, Fujimoto H, Imaizumi K (1989) Size-dependent mortality of red sea bream *pagrus major* juveniles released with fluorescent otolith-tags in News Bay. Jpn J Fish Biol 35(Supplement A):59–69
21. Peterman RM (1991) Density-dependent marine processes in north Pacific salmonids: lessons for experimental design of large scale manipulations of fish stocks. ICES Mar Sci Symp 192:69–77
22. Hilborn R (1999) Confessions of a reformed hatchery basher. Fisheries 24:30–31
23. Kitada S, Taga Y, Kishino H (1992) Effectiveness of a stock enhancement program evaluated by a two-stage sampling survey of commercial landings. Can J Fish Aquat Sci 49:1573–1582
24. Sudo HT, Goto R, Ikemoto MT, Azeta M (1992) Mortality of reared flounder (*Paralichthys olivaceus*) juveniles released in Shijiki Bay. Bull Seikai Natl Fish Res Inst 70:29–37
25. Fujita T, Mizuta T, Nemoto Y (1993) Stocking effectiveness of Japanese flounder *Paralichthys olivaceus* fingerlings released in the coast of Fukushima Prefecture. Saibai Giken 22:67–73
26. Yamashita Y, Nagahora S, Yamada H, Kitagawa D (1994) Effects of release size on survival and growth of Japanese flounder *Paralichthys olivaceous* in coastal waters off Iwate Prefecture, northeastern Japanese. Mar Ecol Prog Ser 105:269–276
27. Svåsand T, Jorstad T, Kristiansen TS (1990) Enhancement studies of coastal cod in western Norway. Part I. Recruitment of wild and reared cod to a local spawning stock. J Cons Intl Expl Mer 47:5–12
28. Svåsand T, Kristiansen TS (1990) Enhancement studies of coastal cod in western Norway. Part II. Migration of reared coastal cod. J Cons Intl Expl Mer 47:13–22
29. Kristiansen TS, Svåsand T (1990) Enhancement studies of coastal cod in western Norway. Part III. Interrelationships between reared and indigenous cod in a nearly land-locked fjord. J Cons Intl Expl Mer 47:23–29
30. Svåsand T, Kristiansen TS (1990) Enhancement studies of coastal cod in western Norway. Part IV. Mortality of reared cod after release. J Cons Intl Expl Mer 47:30–39
31. Nordheide JT, Salvanes AGV (1991) Observations on reared newly released and wild cod (*Gadus morhua* L.) and their potential predators. ICES Mar Sci Symp 192:139–146
32. Leber KM (1995) Significance of fish size-at-release on enhancement of striped mullet fisheries in Hawaii. J World Aquac Soc 26:143–153
33. Leber KM, Brennan NP, Arce SM (1995) Marine enhancement with striped mullet: are hatchery releases replenishing or displacing wild stocks. Am Fish Soc Symp 15:376–387
34. McEachron LW, McCarty CE, Vega RR (1995) Beneficial uses of marine fish hatcheries: enhancement of red drum in Texas coastal waters. Am Fish Soc Symp 15:161–166
35. Kent DB, Drawbridge MA, Ford RF (1995) Accomplishments and roadblocks of a marine stock enhancement program for white seabass in California. Am Fish Soc Symp 15:492–498
36. Willis SA, Falls WW, Dennis CW, Roberts DE, Whitechurch PG (1995) Assessment of effects of season of release and size at release on recapture rates of hatchery-reared red rum (*Sciaenops ocellatus*) in a marine stock enhancement program in Florida. Am Fish Soc Symp 15:354–365
37. Leber KM, Arce SM, Sterritt DA, Brennan NP (1996) Marine stock-enhancement potential in nursery habitats of striped mullet, *Mugil cephalus*, in Hawaii. Fish Bull US 94:452–471
38. Leber KM, Arce SM (1996) Stock enhancement effect in a commercial mullet *Mugil cephalus* fishery in Hawaii. Fish Manag Ecol 3:261–278
39. Leber KM, Blankenship HL, Arce SM, Brennan NP (1997) Influence of release season on size-dependent survival of cultured striped mullet, *Mugil cephalus*, in a Hawaiian estuary. Fish Bull US 95:267–279
40. Rimmer MA, Russell DJ (1998) Survival of stocked barramundi, *Lates calcarifer* (Bloch), in a coastal river system in far northern Queensland, Australia. Bull Mar Sci 62:325–336
41. Lockwood SJ (1991) Stock enhancement. Special session at the ecology and management aspects of extensive mariculture. In: ICES marine science symposia 192, Nantes. International Council for the Exploration of the Sea, Copenhagen
42. WAS (1991) Enhancement of natural fisheries through aquaculture. In: Special session at 22nd annual conference and exposition, San Juan, Puerto Rico. Programs and abstracts. World Aquaculture Society, San Juan
43. AFS (1993) Emerging marine fish enhancement and evaluation. In: Special session at 123rd annual meeting of the American Fisheries Society, Portland. Book of Abstracts
44. EAS (1993) Fisheries and aquaculture interactions. In: Special session at world aquaculture'93, Torremolinos. Abstracts. Special Publication No. 19. European Aquaculture Society, Gent
45. Schramm HL Jr, Piper RG (eds) (1995) Uses and effects of cultured fishes in aquatic ecosystems, vol 15, American fisheries society symposium. American Fisheries Society, Bethesda, 608 pp
46. Travis J, Coleman FC, Grimes CB, Conover D, Bert TM, Tringali M (1998) Critically assessing stock enhancement: an introduction to the Mote symposium. Bull Mar Sci 62(2):305–311

47. Howell BR, Moksness E, Svåsand T (eds) (1999) Stock enhancement and sea ranching. Fishing News Books/Blackwell, Oxford
48. Nakamura Y, McVey JP, Leber KM, Neidig C, Fox S, Churchill K (eds) (2003) Ecology of aquaculture species and enhancement of stocks. In: Proceedings of the thirtieth U.S.-Japan meeting on aquaculture, Sarasota, 3–4 Dec 2001. UJNR Technical Report No. 30
49. Leber KM, Kitada S, Blankenship HL, Svåsand T (eds) (2004) Stock enhancement and sea ranching: developments, pitfalls and opportunities. Blackwell, Oxford, 606 pp
50. Nickum M, Mazik PM, Nickum JG, MacKinlay DD (eds) (2004) Propagated fish in resource management, vol 44, American Fisheries Society symposium. American Fisheries Society, Bethesda, 644 pp
51. Bell JD, Bartley DM, Lorenzen K, Loneragan NR (2006) Restocking and stock enhancement of coastal fisheries: potential, problems and progress. Fish Res 80:1–8
52. Bell JD, Leber KM, Blankenship HL, Loneragan NR, Masuda R (2008) A new era for restocking, stock enhancement and sea ranching of coastal fisheries resources. Rev Fish Sci 16(1–3):1–9
53. Loneragan N, Abraham I (2011) The fourth international symposium on stock enhancement and sea ranching, part of the 9th Asian fisheries and aquaculture forum, Shanghai Ocean University, 21–23 April 2011. Book of Abstracts for Oral and Poster presentations, Shanghai. http://www.SeaRanching4.org/documents/4thISSESR2011.pdf. Accessed Aug 2011
54. Allendorf FW, Phelps SR (1980) Loss of genetic variation in a hatchery stock of cutthroat trout. Trans Am Fish Soc 109:537–543
55. Busac CA, Currens KP (1995) Genetic risks and hazards in hatchery operations: fundamental concepts and issues. Am Fish Soc Symp 15:71–80
56. Leber KM, Blankenship HL (2011) How advances in tagging technology improved progress in a new science: marine stock enhancement. In: McKenzie J, Phelps Q, Kopf R, Mesa M, Parsons B, Seitz A (eds) Advances in fish tagging and marking technology, vol 76, American fisheries society symposium. American Fisheries Society, Bethesda
57. Tringali MD (2006) A Bayesian approach for genetic tracking of cultured and released individuals. Fish Res 77:159–172
58. Kirk R (1987) A history of marine fish culture in Europe and North America. Fishing News Books, Farnham, 192 pp
59. Lorenzen K (2008) Understanding and managing enhancement fisheries systems. Rev Fish Sci 16:10–23
60. Leber KM, Brennan NP, Arce SM (1998) Recruitment patterns of cultured juvenile Pacific threadfin, *Polydactylus sexfilis* (Polynemidae), released along sandy marine shores in Hawaii. Bull Mar Sci 62(2):389–408
61. Leber KM, Cantrell RN, Leung PS (2005) Optimizing cost-effectiveness of size at release in stock enhancement programs. North Am J Fish Manag 25:1596–1608
62. Tringali MD, Leber KM, Halstead WG, McMichael R, O'Hop J, Winner B, Cody R, Young C, Neidig C, Wolfe H, Forstchen A, Barbieri L (2008) Marine stock enhancement in Florida: a multi-disciplinary, stakeholder-supported, accountability-based approach. Rev Fish Sci 16(1–3):51–57
63. Waples RS (1999) Dispelling some myths about hatcheries. Fisheries 26(2):12–21
64. Tringali MD, Leber KM (1999) Genetic considerations during the experimental and expanded phases of snook stock enhancement. Bull Natl Res Inst Aquac Suppl 1:109–119
65. Lorenzen K, Beveridge MCM, Mangel M. Cultured fish: integrative biology and management of domestication and interactions with wild fish. Biol Rev (in press)
66. HRP (2011) US Hatchery Reform Program. http://www.HatcheryReform.us. Accessed Aug 2011
67. Cowx IG (1994) Stocking strategies. Fish Manag Ecol 1:15–31
68. Munro JL, Bell JD (1997) Enhancement of marine fisheries resources. Rev Fish Sci 5:185–222
69. Taylor MD, Palmer PJ, Fielder DS, Suthers IM (2005) Responsible estuarine finfish stock enhancement: an Australian perspective. J Fish Biol 67:299–331
70. Zohar Y, Hines AH, Zmora O, Johnson EG, Lipcius RN, Seitz RD, Eggleston DB, Place AR, Schott EJ, Stubblefield JD, Chung JS (2008) The Chesapeake Bay blue crab (*Callinectes sapidus*): a multidisciplinary approach to responsible stock replenishment. Rev Fish Sci 16:24–34
71. Bartley DM, Kent DB, Drawbridge MA (1995) Conservation of genetic diversity in a white seabass hatchery enhancement program in southern California. Am Fish Soc Symp 15:249–258
72. Masuda R, Tsukamoto K (1998) Stock enhancement in Japan: review and perspective. Bull Mar Sci 62(2):337–358
73. Kitada S (1999) Effectiveness of Japan's stock enhancement programmes: current perspectives. In: Howell BR, Moksness E, Svåsand T (eds) Stock enhancement and sea ranching. Fishing News Books/Blackwell, Oxford, pp 103–131
74. Blaylock RB, Leber KM, Lotz JM, Ziemann DA (2000) The U.S. Gulf of Mexico marine stock enhancement program (USGMSEP): the use of aquaculture technology in "responsible" stock enhancement. Bull Aquac Assoc Can 100:16–22
75. Kuwada H, Masuda R, Kobayashi T, Shiozawa S, Kogane T, Imaizumi K, Tsukamoto K (2000) Effects of fish size, handling stresses and training procedure on the swimming behaviour of hatchery-reared striped jack: implications for stock enhancement. Aquaculture 185:245–256
76. Friedlander AM, Ziemann DA (2003) Impact of hatchery releases on the recreational fishery for Pacific threadfin (*Polydactylus sexfilis*) in Hawaii. Fish Bull 101:32–43
77. Smith TIJ, Jenkins WE, Denson MR, Collins MR (2003) Stock enhancement research with anadromous and marine fishes in South Carolina. In: Nakamura Y, McVey JP, Fox S, Churchill K, Neidig C, Leber K (eds) Ecology of aquaculture species and enhancement of stocks. Proceedings of the thirtieth U.S.–Japan meeting on aquaculture, Sarasota, 3–4 Dec 2001. UJNR Technical Report No. 30. Mote Marine Laboratory, Sarasota, pp 175–190
78. Woodward AG (2003) Red drum stock enhancement in Georgia: a responsible approach. Coastal Resources Division,

Georgia Department of Natural Resources, Brunswick. http://www.peachstatereds.org/approach.pdf. Accessed Oct 2010

79. Jenkins WE, Denson MR, Bridgham CB, Collins MR, Smith TIJ (2004) Year-class component, growth, and movement of juvenile red drum stocked seasonally in a South Carolina estuary. North Am J Fish Manag 24:636–647
80. Kuwada H, Masuda R, Kobayashi T, Kogane T, Miyazaki T, Imaizumi K, Tsukamoto K (2004) Releasing technique in striped jack marine ranching: pre-release acclimation and presence of decoys to improve recapture rates. In: Leber KM, Kitada S, Blankenship HL, Svåsand T (eds) Stock enhancement and Sea ranching: developments, pitfalls and opportunities. Blackwell, Oxford, pp 106–116
81. Fairchild EA, Fleck J, Howell WH (2005) Determining an optimal release site for juvenile winter flounder *Pseudopleuronectes americanus* (Walbaum) in the Great Bay estuary, NH, USA. Aquac Res 36:1374–1383
82. Mobrand LE, Barr J, Blankenship L, Campton DE, Evelyn TTP, Flagg TA, Mahnken CVW, Seeb LW, Seidel PR, Smoker WW (2005) Hatchery reform in Washington State. Fisheries 30:11–23
83. Eggleston DB, Johnson EG, Kellison GT, Plaia GR, Huggett CL (2008) Pilot evaluation of early juvenile blue crab stock enhancement using a replicated BACI design. Rev Fish Sci 16:91–100
84. Gardner C, Van Putten EI (2008) The economic feasibility of translocating rock lobsters in increase yield. Rev Fish Sci 16:154–163
85. Karlsson S, Saillant E, Bumguardner BW, Vega RR, Gold JR (2008) Genetic identification of hatchery-released red drum in Texas bays and estuaries. North Am J Fish Manag 28:1294–1304
86. Le Vay L, Lebata MJH, Walton M, Primavera J, Quinitio E, Lavilla-Pitogo C, Parado-Estepa F, Rodriguez E, Ut VN, Nghia TT, Sorgeloos P, Wille M (2008) Approaches to stock enhancement in mangrove-associated crab fisheries. Rev Fish Sci 16:72–80
87. Potter IC, French DJW, Jenkins GI, Hesp SA, Hall NG, de Lestang S (2008) Comparisons of growth and gonadal development of otolith-stained cultured black bream, *Acanthopagrus butcheri*, in an estuary with those of its wild stock. Rev Fish Sci 16:303–316
88. Purcell SW, Simutoga M (2008) Spatio-temporal and size-dependent variation in the success of releasing cultured sea cucumbers in the wild. Rev Fish Sci 16:204–214
89. Støttrup JG, Overton JL, Paulsen H, Mollmann C, Tomkiewicz J, Pedersen PB, Lauesen P (2008) Rationale for restocking the Eastern Baltic cod stock. Rev Fish Sci 16:58–64
90. Taylor MD, Suthers IM (2008) A predatory impact model and targeted stock enhancement approach for optimal release of mulloway (*Argyrosomus japonicus*). Rev Fish Sci 16:125–134
91. Walters CJ, Hilborn R (1978) Ecological optimization and adaptive management. Ann Rev Ecol Syst 9:157–188
92. Leber KM (2002) Advances in marine stock enhancement: shifting emphasis to theory and accountability. In: Stickney RR, McVey JP (eds) Responsible marine aquaculture. CABI Publishing, New York, pp 79–90
93. Grimes CB (1998) Marine stock enhancement: sound management or techno-arrogance? Fisheries 23(9):18–23
94. Hilborn R (1998) The economic performance of marine stock enhancement projects. Bull Mar Sci 62:661–674
95. Serafy JE, Ault JS, Capo TR, Schultz DR (1999) Red drum, *Sciaenops ocellatus*, stock enhancement in Biscayne Bay, FL, USA: assessment of releasing unmarked early juveniles. Aquac Res 30:737–750
96. Bartley DM, Bell JD (2008) Restocking, stock enhancement, and sea ranching: arenas of progress. Rev Fish Sci 16:357–364
97. Arbuckle M, Metzger M (2000) Food for thought. A brief history of the future of fisheries management. Challenger Scallop Enhancement Company, Nelson
98. Tringali MD, Seyoum S, Wallace EM, Higham M, Taylor RG, Trotter AA, Whittington JA (2008) Limits to the use of contemporary genetic analyses in delineating biological populations for restocking and stock enhancement. Rev Fish Sci 16:111–116
99. Garaway CJ, Arthur RI, Chamsingh B, Homekingkeo P, Lorenzen K, Saengvilaikham B, Sidavong K (2006) A social science perspective on stock enhancement outcomes: lessons learned from inland fisheries in southern LAO PDR. Fish Res 80:37–45
100. Brennan NP, Walters CJ, Leber KM (2008) Manipulations of stocking magnitude: addressing density-dependence in a juvenile cohort of common Snook (*Centropomus undecimalis*). Rev Fish Sci 16:215–227
101. Drummond K (2004) The role of stock enhancement in the management framework for New Zealand's southern scallop fishery. In: Leber KM, Kitada S, Blankenship HL, Svåsand T (eds) Stock enhancement and Sea ranching: developments, pitfalls and opportunities. Blackwell, Oxford, pp 397–411
102. Uki N (2006) Stock enhancement of the Japanese scallop *Patinopecten yessoensis* in Hokkaido. Fish Res 80:62–66
103. Stoner AW (1994) Significance of habitat and stock re-testing for enhancement of natural fisheries: experimental analyses with queen conch *Strombus gigas*. J World Aquac Soc 25:155–165
104. Leber KM (1999) Rationale for an experimental approach to stock enhancement. In: Howell BR, Moksness E, Svasand T (eds) Stock enhancement and sea ranching. Blackwell, Oxford, pp 63–75
105. Agnalt AL, Jørstad KE, Kristiansen T, Nøstvold E, Farestveit E, Næss H, Paulsen LI, Svåsand T (2004) Enhancing the European lobster (*Homarus gammarus*) stock at Kvitsoy Islands: perspectives on rebuilding Norwegian stocks. In: Leber KM, Kitada S, Blankenship HL, Svåsand T (eds) Stock enhancement and sea ranching: developments, pitfalls and opportunities. Blackwell, Oxford, pp 415–426
106. Kitada S, Kishino H (2006) Lessons learned from Japanese marine finfish stock enhancement programs. Fish Res 80:101–112
107. Bartley DM (1999) Marine ranching: a global perspective. In: Howell BR, Moksness E, Svasand T (eds) Stock enhancement and sea ranching. Blackwell, Oxford, pp 79–90

108. Lorenzen K (2006) Population management in fisheries enhancement: gaining key information from release experiments through use of a size-dependent mortality model. Fish Res 80:19–27
109. Medley PAH, Lorenzen K (2006) EnhanceFish: a decision support tool for aquaculture-based fisheries enhancement. Imperial College, London. http://www.aquaticresources.org/enhancefish.html. Accessed Aug 2011
110. Ye Y, Loneragan N, Die DJ, Watson R, Harch B (2005) Bioeconomic modeling and risk assessment of tiger prawn (*Penaeus esculentus*) stock enhancement in Exmouth Gulf, Australia. Fish Res 73:231–249
111. Yamashita Y, Yamada H (1999) Release strategy for Japanese flounder fry in stock enhancement programmes. In: Howell BR, Moksness E, Svasand T (eds) Stock enhancement and sea ranching. Blackwell, Oxford, pp 191–204
112. Tsukamoto K, Kuwada H, Uchida K, Masuda R, Sakakura Y (1999) Fish quality and stocking effectiveness: behavioral approach. In: Howell BR, Moksness E, Svasand T (eds) Stock enhancement and sea ranching. Blackwell, Oxford, pp 205–218
113. Brennan NP, Darcy MC, Leber KM (2006) Predator-free enclosures improve post-release survival of stocked common snook. J Exp Mar Biol Ecol 335:302–311
114. Lipcius RN, Eggleston DB, Schreiber SJ, Seitz RD, Shen J, Sisson M, Stockhausen WT, Wang HV (2008) Importance of metapopulation connectivity to restocking and restoration of marine species. Rev Fish Sci 16:101–110
115. Hervas S, Lorenzen K, Shane MA, Drawbridge MA (2010) Quantitative assessment of a white seabass (*Atractoscion nobilis*) stock enhancement program in California: post-release dispersal, growth and survival. Fish Res 105:237–243
116. Smedstad OM, Salvanes AGV, Fosså JH, Nordeide JT (1994) Enhancement of cod, *Gadus morhua* L., in Masfjorden: an overview. Aquac Fish Manag 25:117–128
117. Otterå H, Kristiansen TS, Svåsand T, Nødtvedt M, Borge A (1999) Sea ranching of Atlantic cod (*Gadus morhua* L.): effects of release strategy on survival. In: Howell BR, Moksness E, Svåsand T (eds) Stock enhancement and sea ranching. Fishing News Books/Blackwell, Oxford, pp 293–305
118. Wang Q, Wu H, Liu H, Wang S (2011) Ecosystem based sea ranching in Zhangzidao in northern yellow sea. In: Fourth international symposium on stock enhancement and sea ranching, Shanghai. Abstract, available within pdf file. http://www.SeaRanching4.org/documents/4thISSESR2011.pdf. Accessed Aug 2011
119. Becker P, Barringer C, Marelli DC (2008) Thirty years of sea ranching Manila clams (*Venerupis philippinarum*): successful techniques and lessons learned. Rev Fish Sci 16:44–50
120. Chaplin J, Hesp A, Gardner M, Cottingham A, Phillips N, Potter I, Jenkins G (2011) Biological performance and genetics of restocked and wild black sea bream in an Australian estuary. In: Fourth international symposium on stock enhancement and sea ranching, Shanghai. Abstract, available within pdf file. http://www.SeaRanching4.org/documents/4thISSESR2011.pdf. Accessed Aug 2011
121. Jenkins WE, Smith TIJ, Denson MR (2004) Stocking red drum: lessons learned. Am Fish Soc Symp 44:45–56
122. Miller JM, Walters CJ (2004) Experimental ecological tests with stocked marine fish. In: Leber KM, Kitada S, Blankenship HL, Svåsand T (eds) Stock enhancement and sea ranching: developments, pitfalls and opportunities. Blackwell, Oxford, pp 142–152
123. HSRG (2011) Hatchery scientific review group, puget sound and coastal Washington hatchery reform project: applying the principles of reform to Western Washington's hatcheries. http://www.lltk.org/improving-management/hatchery-reform/hrp/hsrg. Accessed Aug 2011
124. Loneragan N, Jenkins G, Taylor M (2011) Stock enhancement and restocking in Australia and opportunities for finfish, particularly in Western Australia. In: Fourth international symposium on stock enhancement and sea ranching, Shanghai. Abstract, available within pdf file. http://www.SeaRanching4.org/documents/4thISSESR2011.pdf. Accessed Aug 2011
125. ISSESR (2011) The international symposium on stock enhancement and sea ranching, Shanghai. http://www.SeaRanching.org. Accessed Aug 2011

Marine Life Associated with Offshore Drilling, Pipelines, and Platforms

Martin Hovland
Centre for Geobiology, University of Bergen, Bergen, Norway
Statoil ASA, Stavanger, Norway

Article Outline

Glossary

OHI The Offshore Hydrocarbon Industry (OHI) searches for natural accumulations (reservoirs) of oil and gas (hydrocarbons) and develops the means to extract and distribute (transport) them.

Platform An artificial structure designed to drill for hydrocarbons and/or produce (extract) and distribute hydrocarbons offshore in water depths up to 3 km. A platform can either be floating, semi-submersible, or fixed to the seafloor.

Subsea template A structure normally constructed of steel tubing, designed for a variety of purposes within the OHI. A normal subsea production template has up to four wellheads and has typical dimensions of 20 m × 20 m × 10 m.

ROV Remotely Operated Vehicle (ROV) is a remotely controlled underwater vehicle of variable size (from <1m long, up to about 3 m in length). The vehicle is normally fitted with propellers (thrusters), lights, cameras, manipulator arms, and other sensors and devices depending on its operational task.

Trunk pipeline A pipeline designed to transport large quantities of natural gas or oil over long distances (up to 1000 km). Normally, they have diameters between 20″ and 44″ (inner diameter of the steel pipe). Before laid on the seafloor, they are coated with varying thicknesses of concrete coating for added weight, to prevent them from becoming buoyant.

Umbilical A specially designed flexible, multipurpose cable used for powering underwater equipment (including ROVs and subsea templates) and also used for sending and receiving control and sensor signals. Umbilicals can contain combinations of electrical cables and optical fibers.

Cold-water coral reef A mounded natural structure on the seafloor consisting of live animals and dead remains and sediments. The mound is partly constructed by colonizing corals that are not dependent on sunlight (i.e., ahermatypic corals) such as the most common species: *Lophelia pertusa*.

Cold seep A location on the seafloor where natural fluids (gas and liquids) seep upward from the substratum, into the overlying water column.

Iceberg ploughmark Up to 100 m wide and many kilometer long furrows in the seafloor, produced by the action of drifting grounded icebergs. Off Mid- and Northern-Norway and several other places such (relict) furrows remain from the last glaciation.

Fish sighting The underwater visual detection (recording) of fish (here, larger than 0.5 m in length) using submersible vehicles with lights and cameras, such as ROVs.

Definition of the Subject

The offshore hydrocarbon, "oil," industry (OHI) searches for oil and natural gas by drilling exploration wells as deep as 10 km below the seafloor. When a commercial oil or gas field has been documented by such drilling, the exploitation of the resource will start by developing the field and the construction of production units and transportation infrastructure. Until only 15 years ago, this meant the construction of large, concrete-based or steel "jacket" production platforms. Because of intense research and technological development, many of the new offshore hydrocarbon fields are developed with smaller remotely controlled subsea steel structures placed directly on the seafloor, often without any infrastructure visible above the water. These new fields are produced remotely over distances of up to 150 km, with fiber optical cables, satellite communication, umbilicals, and pipelines.

In contrast to the other main (traditional) offshore industry, for example, the fishery industry, the OHI has employed strict environmental rules and regulations, which are efficiently practiced in most countries. These ensure little harm to sensitive marine organisms during normal field development and production. In addition to obeying the imposed rules and regulations, the OHI is, by tradition, constantly developing new and more cost-effective and environmentally friendly technology and infrastructure.

With knowledge and experience from 30 years of underwater detailed mapping and visual observations, mainly from the North Sea, spanning from predrilling seafloor surveys to annual surveys of pipelines and platforms, it is found that the marine life (the visual mega-fauna, at least) apparently benefits from the OHI-related installations on the seafloor. The reason being improved shelter conditions for large fish and also for spawning fish, and also an increased amount of energy (nutrients and seston) available in the water mass near these human-made structures, some of which act as artificial reefs [1]. The future needs for improved management of the marine biological resources, including the valuable deepwater corals and natural fish stocks, can be done by increased awareness of underwater life in general, via live video footage released to the public by, for example, the OHI. Furthermore, it also calls for academic scientific research into how best visual documentation of the seafloor can be used for an improved understanding of the complex underwater ecology and biodiversity change.

Introduction

In an increasingly energy hungry society, the quest for finding and exploiting underground oil and gas (hydrocarbon) resources is being continuously improved. Whereas the world's total onshore hydrocarbon production is gradually decreasing, the OHI is currently increasing its production volume. According to Maurer [2], ecological systems are complex and combine both idiosyncratic and unpredictable outcomes with strong constraints on system structure that makes them paradoxically both deterministic and unpredictable at the same time. Because of this, there has been no universal theory to guide research on ecological phenomena.

Over the last 50 years, the OHI has, unfortunately, inflicted several enormous oil-spills on the marine and coastal environments. There have at least been five such episodes that should never have occurred: The blowout and spill in the Santa Barbara basin, off Los Angeles (January, 1969), the Ekofisk Bravo blowout in the Norwegian-sector of the North Sea (April, 1977), the Ixtoc blowout, Mexican-sector of the Gulf of Mexico (GoM) (June, 1979), the Piper Alpha disaster, UK-sector of the North Sea (July, 1988), and lastly, but not least, the Deepwater Horizon blowout and disaster, US-sector of the GoM (May, 2010). Apart from these unfortunate, generally short-lived (less than 2 years), environmental inflictions, the OHI at large appears to be environmentally friendly, as will be discussed herein. This notion has been documented by extensive seafloor mapping and annual visual inspections of platforms, pipelines, and other infrastructure. Thus, rather than representing a threat to marine life in general, the OHI is, at least in the North Sea, a benefit to marine life in general. This is not only because, by its design, it protects numerous fish against industry fishing and trawling, but also because the large artificial underwater steel and concrete constructions represent geometrically complex structures in an otherwise mostly structureless seafloor environment. Furthermore, the industry is continuously improving its methods for underwater mapping, inspection, and monitoring of the environment.

This assessment of marine life associated with normal offshore drilling, pipelines, and platforms stems from over 30 years of unique visual observation by manned submarines (1977–1981) and ROVs (remotely operated vehicles), (1979–2010). It is based on the active participation and responsibility for conducting detailed mapping surveys of the seafloor, visual documentation, coupled with remotely sensed (geophysical data). The current experience covers large expanses of virgin seafloor, stretching from the Shtokman field at 73.6°N, in the eastern Barents Sea, south to 51°N, off Dunkerque, France. A total of 522 fixed production-related structures (platforms and subsea templates) have been installed on the Norwegian Continental Shelf, and over 7,000 km of trunk pipelines have been constructed in these regions during this time-span. Thus, there is a unique variety of first-hand specific knowledge that can be shared from the numerous site surveys of platform locations, pre-lay visual surveys (conducted before the laying of the long, trunk pipelines) of the seafloor, to annual inspections of the constructed pipelines.

However, the main difficulty is how to describe and disseminate this unique visual OHI-related underwater experience and information in a way that can be used by marine scientists in a quantitative manner. This task is envisioned to resemble that faced by the pioneering land-explorers after their long treks across previously unknown parts of the globe, during the "age of discovery," a couple of centuries ago. The narration, therefore, will be fragmentary, as most of what is observed on the seafloor is new, and as most of the water and seafloor bordering onto the visually observed space is virtually unknown, despite it occurring in some of the world's most fished and scientifically studied oceanic regions (the North Atlantic Ocean).

The Offshore Hydrocarbon Industry (OHI) and the "Second Surface"

A Brief History of the OHI

The onshore hydrocarbon industry started moving out into shallow waters sometime in the early 1930s, offshore Venezuela (Lake Maracaibo), offshore the states of California and Louisiana, USA, and in the Caspian Sea, offshore Baku, Azerbaijan. The first installations were simple steel and wood constructions built in knee-deep waters. However, their size and complexity was gradually increased with increasing water depth, up to several tens of meters. Simple steel jacket drilling towers were constructed and there were bridges and roads built on piled steel and wood foundations, often in a hap-hazard manner. After sometime, there were many accidents and mishaps, before improvements were made and special standards were invoked. The one single event that hit the OHI and aroused the world's environmental conscience was the big blowout oil spill in Santa Barbara, offshore Los Angeles, California on January 29, 1969. This also had immense consequences for the stricter regulations imposed on offshore drilling and the exploitation of offshore oil and gas. Even though no people were killed, this event made such a graphic impression on the population of southern California that in the following spring, "Earth Day" was born. Many consider the publicity surrounding the oil spill a major impetus to the environmental movement.

In Europe, the OHI started with the development of the UK southern gas fields off the east coast of England, in the mid-1960s. Here, the platforms and pipelines met the tough environment of the North Atlantic. New rules and regulations, British North Sea Standards were imposed. In 1967, the OHI moved even further north, in the North Sea, to Norwegian waters. The Norwegian Petroleum Directory (NPD) and The Norwegian state oil company, Statoil, were born, some years later. Although the giant oil field Ekofisk was developed with similar standards as in the UK southern gas fields and the Forties and Piper fields of the mid-UK North Sea, the Norwegian fields still further north, such as Statfjord, Gullfaks, and Troll, had to withstand even tougher environmental conditions. These fields, located at water depths between 130 and 320 m were therefore developed with giant concrete "gravity base" platforms, as the underwater technology evolution was not ready for moving delicate equipment like pumps, electronics, and gauges under water. When the Troll A concrete platform was towed out from Stavanger, and placed on the seafloor over the giant Troll field on May 17, 1995, it became the largest human-made structure ever to be moved. It measures a total of 472 m in height, from the top of the drilling tower to the bottom of the concrete skirts that penetrate 30 m into the soft clays at the location where it is still producing oil and gas, off Bergen, Norway. The platform houses about 200 workers who stay on board for 14 days at a time, rotating in and out by helicopter.

The rotation occurs all year round, even during the darkest and stormiest winter months (December through February). This platform produces about 18% of the total gas consumption of Germany and is, therefore, of immense value both for the owners (Statoil, Shell, and the Norwegian government) and for the consumers in Germany and surrounding countries. The gas is transported through 36″ (36 in.) and 40″ concrete-coated steel pipelines, welded together on board huge offshore pipe-laying vessels and placed carefully onto the seafloor. Such huge "trunk"-pipelines criss-cross from the Norwegian fields to processing plants on mainland Norway, and are re-routed from there, through other gas export pipelines to England, Germany, Belgium, and France. The largest pipe, the Langeled pipeline, was constructed between 2004 and 2006. It is a 40″ and 44″ diameter pipeline of 1,200 km length, which originates from the onshore processing plant at Nyhavna, south of Trondheim. From there, it runs south to the Sleipner field in the middle of the North Sea, and continues to Immingham on the east coast of the UK.

During the last 15 years, subsea technology has developed fast and most modern fields are constructed solely with remote-controlled subsea structures. The Snehvit field in the Barents Sea is, for example, produced through three subsea templates (steel structures) with several production wells in each template. The field lies 135 km from shore and is remotely operated from the onshore production plant at Melkøya near Hammerfest, the world's northernmost city. At present, Norway is the third world's largest exporter of crude oil, and it runs 522 offshore fixed production-related structures of which 365 are subsea and fully submerged. In year 2009, Norway also exported a total volume of 96.6 billion Standard cubic meters (Bn Sm^3) of natural gas to Europe through the trunk pipelines.

The Second Surface of Earth

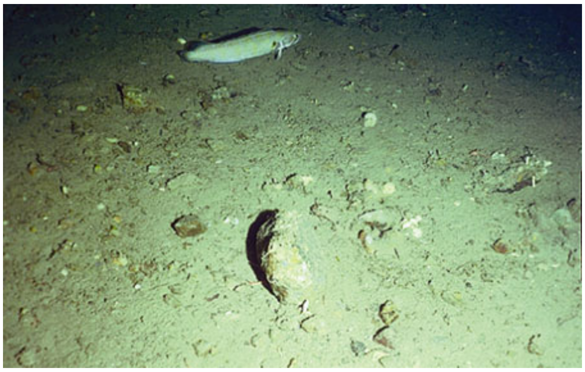
The seafloor is the "Second Surface" of Earth, indicating that it is hidden in many ways. It covers an area which is about three times larger than the visible land surface. Most of this surface is still unknown – because water is a "black body" substance when it comes to the electromagnetic spectrum. It absorbs most of the visual light that encounters it. Therefore, in contrast to sound waves, the light rays have very low transit ranges through water. The photic zone of the ocean, into which solar rays can penetrate, are reckoned to be down to a maximum of just over 100 m in the clearest waters, that is, water with little seston and other particles. This is not very deep, considering the mean depth of the ocean being about 3,500 m. Because of this lack of efficient visual access, the Second Surface is only beginning to be explored in detail. This surface ranges in depth, from 0 m at the coastlines, to about 300 m on the continental shelves, then down to 5,000 m (5 km) on the great abyssal plains, to over 11 km in the deepest trenches. Because most of Earth's surface is covered by water, the Second Surface represents a very significant and essentially important entity. So far, remotely sensed (acoustic) surveys only cover about 10% of it (i.e., indirectly, with data that needs interpretation by geophysicists) and only less than 1% visually (i.e., directly with cameras). This means there exists more visual documentation of the surfaces of both Moon and Mars, than of the immensely more important Second Surface of Earth. However, from sediment sampling, fishing (trawling), scientific scraping (dredge sampling), and drilling, in all oceans, over time, it is currently known that the Second Surface mostly consists of mud (clay), sand, rock, and in some areas metals and salts. But, because being flooded by water, it is both pressurized, and buoyed at the same time, and behaves accordingly, which is often totally different to the well-known land surface. Furthermore, on average, the Second Surface has a much thinner crust than the onshore continental crust and is more likely to be exposed to high heat flow from the Earth's interior. Along the tectonic plate boundaries (mid-ocean spreading zones and subduction zones), the high heat flow induces underground convection fluid currents and the venting of warm fluids in hydrothermal vent systems [3].

Although academic research institutions and consortia, such as the Integrated Ocean Drilling Program (IODP), perform many types of investigations at water depths to about 6 km, the mining industry is working at depths to about 5 km, and the fishing fleet is gradually trawling to depths greater than 2 km, the OHI is currently working to water depths of about 3 km. However, about 90% of its activity still occurs at water depths between 100 m and 1,500 m. These are the water depths, therefore, which will be addressed herein. From the extensive mapping-, construction-, and visual inspection-work performed by OHI at these depths, several places in the world, there is some general and also specific knowledge about processes and marine life that can be disseminated, including some new discoveries.

Whereas the fishing industry tends to operate in a "blind" mode when it comes to the seafloor, the OHI naturally operates with more caution, partly in order of preventing damage to sensitive and costly equipment and structures, and partly because of law enforcement (at least in US and European, including Norwegian waters). Thus, no drill-site is drilled without a proper predrilling assessment of the seafloor, whereby any significant physical obstructions and known sensitive organisms, including chemosynthetic fauna and coral reefs are documented beforehand. The problem with the bottom trawling of the fishery industry is the insensitivity to what is down there, that is, "indiscriminate obliteration" on the seafloor. As long as such bottom trawling is legal practice, all sessile organisms in the world are actually threatened by trawl-board disturbance, at least those living at water depths shallower than 2,000 m water depth.

The General Background Seafloor Life

In order to set the OHI-related marine fauna observations into perspective, the general background seafloor has to be described. In the depth interval 100–1,500 m, the general background seafloor (about 90% of the total area within this depth range) is drab and appears relatively "uninteresting" (Fig. 1), just like the enormous sand fields of a desert on land. However, in most areas, the drabness is spotted with small hubs of life, like oases in the same desert. There are vast areas with level, muddy bottoms. Several studies have shown,



Marine Life Associated with Offshore Drilling, Pipelines, and Platforms. Figure 1
The general seafloor off Mid-Norway looks like this. In the background is a Tusk (*Brosme brosme*). The stones (cobbles and gravel) have been washed out of the underlying clay-dominated till. Some of the larger stones can be seen to be colonized by invertebrates of different colors

Marine Life Associated with Offshore Drilling, Pipelines, and Platforms. Figure 2
A large erratic boulder located inside a pockmark crater at 280 m water depth off the Island of Fugløy off Northern-Norway, north of the Polar Circle. The organisms seen colonizing the boulder are mainly filterfeeders, sea anemones, and serpulid tube worms

however, that, for the deep-sea biota, there is a distinct decrease in population differentiation and species diversity with depth [4, 5]. Any large erratic boulder, rock outcrop, or wreck is colonized by invertebrates, which seem to attract also other marine life, including fish (Fig. 2).

Piepenburg and co-workers [6] used classical marine biological methods to study patterns and determinants of the distribution and structure of benthic faunal assemblages in the northern North Atlantic. Using a suite of sampling methods including corers, trawls, and seabed imaging (benthic spot photography), they managed to "adequately probe various benthic community fractions, such as foraminifers, poriferans, macrobenthic endofauna, peracarid crustaceans, and megabenthic epifauna." The general patterns they found were, not unexpectedly, a depth zonation, and also a significant decline in biomass and abundance by as much as two and three orders of magnitude. These were the most conspicuous general patterns detected. However, in terms of species richness, no common trend for water depth or latitude was perceivable. They especially studied the East Greenland continental shelf margin between 68°N and 81°N at water depths between 40 and 3,700 m. Here they found relatively productive hydrographic zones being the marginal ice zones, polynyas, and anti-cyclonic gyres. They interpret this as being evidence for the importance of water column processes for subsequent food availability being the major determinants for the benthic assemblages and the significance of pelago-benthic coupling in the study in general [6]. This is not surprising, food availability being the most necessary ingredient for life in general. When it came to the distribution of megafaunal species, such as echinoderms, it was found that community patterns on a 10 km scale and the dispersion of organisms on a 100 m scale were best explained by the seafloor properties. This means that macrofauna is dependent on structure and type of the seafloor sediments and topography. Furthermore, they found no evidence for a direct pelago-benthic coupling, irrespective of water depths. These contrasting findings emphasize that the relative importance of potential community determinants can change with both spatial scale and life traits, for example, body size, mobility and feeding ecology, of organisms considered [6]. Thus, the stage is set for a narration of discoveries made by the OHI and the associated biological and physical research performed for this industry by academic researchers and institutions during the period 1980–1998. After 1998, the academic institutions have picked up many of the research leads pioneered by OHI-activity within marine geophysics and biology.

Unique Processes and Biotypes Initially Studied due to OHI-Activity

Since the late 1960s, new processes and features have been discovered on the Second Surface. Some of them are completely unique to the underwater world at water depths of 100–1,500 m. Perhaps one of the most surprising revelations is that this type of mapping and areal seafloor documentation became instrumental for the discovery of several previously unknown natural conditions of the seafloor. Thus, at least three unique discoveries were made as a result of such surveys: (1) A biological seepage relationship, (2) The discovery of pockmark craters and their potential significance for marine life, and (3) The discovery of myriads of large, cold-water coral reefs (also called deepwater coral reefs). Although the results were not possible without cooperation with academic institutions, especially in the UK, USA, Germany, France, and Norway, the pioneering discoveries were instigated due to OHI-activities, mainly in the Gulf of Mexico and the North Sea. The discoveries are mentioned here, as they sometimes are relevant to the marine life observed on the seafloor. These processes and features are:

1. Venting of reduced organic fluids (fluid flow, or "cold seeps")
2. Crater formation, by fluid flow
3. Bioherms, including the cold-water coral reefs

One of the recent wide-scope books on benthic life in the North Atlantic [7], actually fails to mention the existence of prolific deepwater coral reefs occurring there. In the book, which is aptly titled: "The Northern North Atlantic – A Changing Environment" neither deepwater coral reefs, *Lophelia pertusa*, nor "*Lophelia*-reefs" are found in the index, or at all mentioned in the text. Why is it that thousands of large coral reefs, some known to science for at least 200 years and from the early 1990s published by OHI-related scientists, manage to avoid mention (recognition) in such apparently authoritative scientific literature? Could it be that only "Classical marine biological" results are recognized? The publisher claims that: "the Greenland-Iceland-Norwegian Seas can now be considered one of the best studied subbasins of the world's oceans" [7]. But, even so, the information published in this book is important, as it provides the necessary background knowledge about life in general on the seafloor "desert," outside the coral reefs.

Venting of Reduced Organic Fluids Generally, the ocean floor is covered in thick sediments that deposit by gravitation, with particles sinking through the water column and accumulating in thick layers on the Second Surface. The fluids, including petroleum gas and liquids (hydrocarbons) trapped underneath such sediments are lighter than the solids and, therefore, move upward to surface at discrete locations due to buoyancy. This process is also called "migration" and where the flow penetrates the Second Surface from below, it is called marine fluid flow [8]. The discrete locations where the fluids occur at the surface are called "cold seep" locations. Depending on the geological setting, the distance between each cold seep location on the seafloor varies considerably, from kilometers or miles, to only several meters. However, cold seeps are important for life within, on, and above the Second Surface because they represent transport pathways for dissolved chemical constituents and sustain unique oasis-type ecosystems at the seafloor [9]. Fluids expelled through seeps contain re-mineralized nutrients (silica, phosphate, ammonia, and alkalinity) and hydrogen sulfide, as well as dissolved and free methane from microbial degradation of sedimentary organic matter. Because methane gas molecules (CH_4) have the highest relative hydrogen content (four hydrogen atoms to one carbon atom) of any organic compound, it represents a valuable energy source to certain primary producers: archaea and bacteria, that is, the methanotrophs and the methane oxidizers. Apart from near-cold seep locations, seawater has generally very low concentrations of methane and other light hydrocarbons, such as ethane (C_2H_6), propane (C_3H_8), butane (C_4H_{10}), and pentane (C_5H_{12}). Perhaps the single most important reaction associated with cold seeps is the anoxic oxidation of methane (AOM) by archaea and sulfate reducing bacteria (SRB), with secondary reactions involving the precipitation of carbonate ($CaCO_3$), in the form of inorganic aragonite and calcite [9]. The OHI has long been interested in these seafloor processes, for various reasons, not least because they contain tell-tale indications of where deep-seated hydrocarbons (reservoirs) may be found.

During a predrilling geophysical site survey, in 1977, at the Tommeliten field in the central North Sea at 78 m water depth, side scan sonar data showed numerous bubble streams emanating from the seafloor, immediately above a buried salt dome [10]. Subsequently, this location was investigated with ROV by Statoil in 1983. The gases leaking naturally through the seafloor were documented to be the reduced organic light hydrocarbon gases (methane to pentane) which also continually charged the upper, porous sediments. An intimate relationship was found between organisms, such as anthozoans and the visual bubbling of gas through the sediments [11]. This visual inspection and sampling of the naturally leaking gas produced several interesting results: (a) documenting small "reefs" or "bioherms" consisting of many kinds of filter-feeders, scavengers, and predators occurring adjacent to the seeps; (b) small depressions (so-called eyed-pockmarks), in the seafloor, with high-density macrofaunal communities in their centers, and (c) white patches of bacterial mats also occurring over relatively large seafloor areas, where the sediments were charged with gas seeping up from deeper layers.

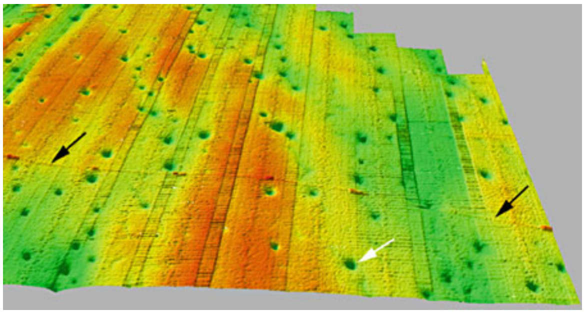
Later studies of this Tommeliten site also showed two other important aspects, relevant to marine life: (1) that the bacterial mats were easily torn apart and carried up into the water column by slight disturbances of the near-bottom water by the ROV and (2) that the seafloor had been partly cemented by methane derived authigenic carbonate rock [8, 11] both within the eyed-pockmarks and elsewhere on the otherwise flat seafloor, near the bacterial mats [12]. Academic research at this active seepage site and also at one similar site near the Gullfaks field, further north, identified a microbial community dominated by sulfur-oxidizing and sulfate-reducing bacteria (SRB) as well as methanotrophic bacteria and archaea. Stable carbon isotope values of specific, microbial fatty acids and alcohols from both the Tommeliten and Gullfaks sites were found to be highly depleted in the heavy isotope ^{13}C, indicating that the microbial community readily incorporated seeping methane or its metabolites [13].

At Tommeliten and Gullfaks there is, therefore, no doubt that the dense bioaccumulations on the seafloor, including, the bioherms, are a direct result of seeping gas (energy and nutrients) from deep below the seafloor. This may also be one of the reasons why there is plenty fish and heavy trawling activity at the Gullfaks location [14]. Although most marine ecologists, environmental scientists, and biologists are used to assess traditional phytoplankton concentrations dependent only on top-down linkages in the food chain, the modern seep studies find more and more bottom-up links [13, 15].

Mud volcanoes are locations where fluids and solids (water, mud, gas, and petroleum) well up through the Earth's surface, driven by overpressures in the subsurface. They both occur on the land surface, as in Azerbaijan [16] and many places on the Second Surface [17, 18]. From studies and surveys at the underwater Håkon Mosby Mud Volcano (HMMV) located at the boundary between the Barents Sea and the Norwegian-Greenland Sea, a relatively prolific and complex ecology has been found. Although only a minor portion of this ecosystem relies directly upon chemosynthetic energy, this portion is probably very important for the sustenance of the system. A simplified food chain for HMMV was published by Vogt et al. [19]: The primary producers are suspected to be methanotrophic bacteria and anoxic methane oxidizers (archae), besides other "conventional" microorganisms relying on added heat and continuous sediment disturbance by the turbid flow of mud from the mud volcanic vent. The secondary consumers were made up of benthic suspension feeders and deposit feeders, such as asteroids, holothurians, etc. In addition, there was a chemosynthetically based food web, relying on symbionts, such as the pogonophoran tube worms, that host endogenic chemosynthetic bacteria [10, 19]. The primary consumers consist of filter-feeders and other predatory invertebrates and vertebrates, such as sponges, crinoids, pycnogonids, basket-stars, and fish, mainly consisting of Eelpout and Skate (Raja spp). A maximum density of one Eelpout for every square meter was documented above portions of the mud volcano [19]. This example also demonstrates the very important influence of substances originating from below the seafloor in modifying and "fertilizing" the immediate seafloor environment, to result in enhanced productivity.

Seabed Pockmarks: Craters Formed by Focused Fluid Flow On the land surface it is often very difficult to detect surface degassing (seepage) events, apart from those associated with water and mud flows, such as at mud volcanoes [20]. On the sediment-covered Second Surface and in some lakes, however, the situation is different. In the late 1960s, numerous craters were found off Nova Scotia. They were called "pockmarks" by their discoverers, Lew King and Brian MacLean of the Bedford Institute of Oceanography (BIO), Canada [21]. Today, it is known that pockmarks occur in certain portions of the seafloor, the world over, and even in some lakes [8, 10]. Also a very close relationship between the pockmarks and local increase in visible seston (including plankton) was found [11]. Although pockmarks ranging in size from the small, "unit-pockmarks" (<5 m diameter, 1 m depth), to normal-pockmarks, complex-pockmarks, and giant-pockmarks (up to 500 m in diameter and 30 m depth), have been found, very little is known about their formation and sustaining mechanisms [8, 22]. Although they are known to be formed as a consequence of buried gas reservoirs and fluid flow, there are only very few pockmarks known to be producing continuous visible bubble streams [23].

After having noticed the marine life occurring inside some of the pockmark explored with ROVs in the North Sea [10], it was recognized that the "classic" (conventional) science of marine geology would probably never have discovered pockmark craters (Fig. 3). The reason being, that this science relied on acquiring numerous spot samples of the seafloor using corers and grab samplers, and determining the nature of the seafloor mainly by measuring the grain-size of the spot sampled sediment grains. Thus, if only one seafloor sample was acquired at 1 km spacing in the Norwegian Trough of the North Sea, where there is an average of 15 pockmarks per square km, it is likely that only one sample out of 100 would happen to sample inside a pockmark crater. This sample would probably turn out to be anomalous (i.e., containing unexpected biota and sediments). Thus, the chances of this sample (one-in-a-hundred) being discarded or recorded as an "anomaly" (or only of "curiosity value") are great. At least it would not turn out in the statistics of the survey, and the chances of it being taken seriously are therefore meager.

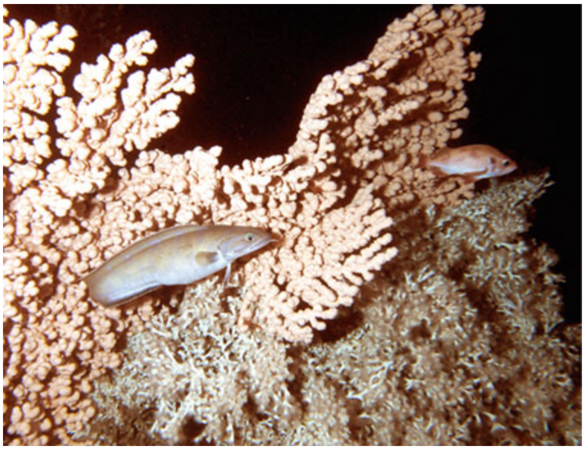


Marine Life Associated with Offshore Drilling, Pipelines, and Platforms. Figure 3
In the 300 m deep Norwegian Trough area of the northern North Sea, there are thousands of normal-pockmarks. They are shown here on a perspective image of a seafloor relief map, with artificial vertical enhancement (x5). The size of the largest pockmark craters is 100 and 8 m depth. The *white arrow* points to a normal-pockmark of about 70 m diameter. The *black arrows* point at a trunk pipeline of 30″ diameter. Lines running north south represent noise in the digital data set, due to inaccuracies in sound velocity and tidal correction

From modern seafloor remote sensing with multi-beam (swathe) echo-sounders to long-range side scan sonars, it is known that the seafloor can be covered in a high density of pockmark craters. The density ranges from zero to more than 20 pockmarks per square km of mapped seafloor (Fig. 3). Because recent studies conclude that pockmarks form due to focused fluid flow [22, 23], the seafloor can generally be divided into "hydraulically active" and "passive" areas. Thus, the active seafloor will have seep manifestations, like pockmark craters and the passive areas will be devoid of craters.

There are two important corollaries to this "hydraulic theory": (a) It is valid for all volumes of soil, which have a porosity system partly filled by liquid and partly filled by gas. It is, therefore, also valid for all ocean depths, lakes and swamps, as the driving gas-type (methane, carbon dioxide, hydrogen sulfide, or hydrogen) is immaterial. (b) Because seeping fluids through the seafloor can be regarded as enhanced energy input to the marine fauna (primary producers especially), the enhanced hydraulic activity manifest by seeps and high density of pockmarks indicates the likelihood of enhanced marine biological productivity.

Cold-Water Coral Reefs of the North Atlantic It is known that the presence of solitary corals, sea pens, sea lilies, and sponges on the deep-sea floor offers rare, firm substrates for sessile organisms in an otherwise generally featureless environment. The relative importance of such biotic habitats for the local biodiversity may, therefore, be greater for the deep-sea than for shallower regions [24]. Those modern reef structures that seem to defy all normal reasoning with respect to location and environment are the deepwater coral and cold-water coral reefs (also named "ahermatypic colonial scleractinians"). In the North Atlantic and Gulf of Mexico, the most common types are those built by the *Lophelia* sp. stony coral, found on both sides of the Atlantic Ocean, including the Reykjanes Ridge, south of Iceland (Fig. 4). Because they are also found as far north as the Polar Circle, in the Barents Sea and off Mid- and Northern-Norway, they must somehow be independent of seasonal sunlight variations. Some of these reefs, found off Ireland and on parts of the Norwegian Continental Shelf had been known to fishermen and biologists for over 200 years [25–30]. Even so, the visual documentation by OHI-surveys during the 1980s and mid-1990s of thousands of large, deepwater coral reefs was completely unexpected to most marine biologists [31–36].

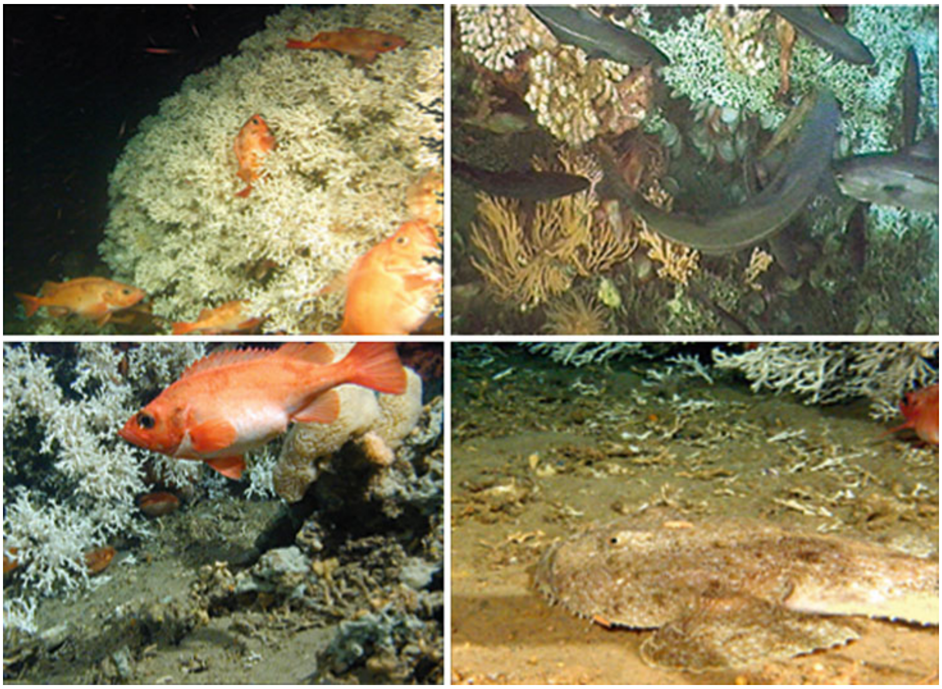


Marine Life Associated with Offshore Drilling, Pipelines, and Platforms. Figure 4
Photograph from a typical Norwegian deep-water coral reef showing two types of corals: the soft-branched octocoral *Paragorgia arborea* (*closest*) and the reef-forming ahermatypic stony coral *Lophelia pertusa*, (*white and pink*). The two most common reef-related fish are also shown here, the Tusk (*Brosme* sp) and the Redfish (*Sebastes* sp) [47]. The height of this coral reef is about 3 m above the surrounding, undisturbed, even seafloor

Their modern rediscovery with broad-scale geophysical mapping and detailed video footage and photographs became a major eye-opener – a revelation that caused an improved reassessment of the complex and dynamic biological production of the North Atlantic Ocean. This modern documentation actually started in June, 1982, with an OHI-related detailed investigation for a potential pipeline route from the Askeladden field in the Barents Sea to Norway. During which, a 15 m high and 50 m wide *Lophelia*-reef was found and visually documented by Statoil [31, 33]. Further pipeline route surveys off Mid-Norway between 1985 and 1990 documented hundreds of similar reefs, until then unknown to science. Until June, 1991, when Statoil invited researchers from five Scandinavian academic and Norwegian authorities to view their unique data set, consisting of detailed geophysical and photographic/video results, these biotic structures had mainly been treated as curiosities and oddities on the seafloor [37]. But, even after this seminar, it was not the Norwegian scientists who first managed to mobilize an academic detailed visual study of the reefs, but the Germans. Freiwald et al. [34] had previously been studying carbonate secreting algae along the northern Norwegian coast. Having learnt from Statoil where the reefs occurred, they targeted them during their next planned cruise, already in 1992 [34]. The main surge of academic interest in the reefs of the North Atlantic came immediately after the publication of large carbonate reefs off Southwest Ireland. This was also partly based on OHI-related exploration geophysical data [33, 35, 38]. Subsequently, the first International conference on cold-water corals was staged in Canada, 2001 [39].

One of the main and controversial questions still remaining to be answered with respect to these impressive biological structures is why they occur in deep and cold water, even north of the Polar Circle, where there is hardly any photosynthesis occurring during the winter months. Even though there is ample evidence suggesting that they rely on extra nutrients and energy originating from below ground, that is, the "hydraulic theory" [11, 31, 35, 36, 40, 41], modern marine biological research has still not found sufficient evidence to support this theory. The main theory of the modern marine ecologists is, according to Buhl-Mortensen et al. [24], that they exist in high latitudes and deep waters because this is where the right water masses occur: ". . .it is not the deep water par *se*, but the distribution of intermediate and deep water masses that controls the bathymetric distribution of these corals. Corals typically create habitats reaching from decimetres to meters above the surrounding seabed and occur on mixed bottoms in areas with relatively high currents." [24], and the argumentation continues: "Colonial scleractinians need hard substrate for settlement. This substrate can be a shell or a pebble, and as soon as one colony is present it provides new hard substrate for subsequent colonisation" [24]. So, the question remains as to why only less than 0.1 per mille (‰) of the total area in the depth zones where they occur is covered by cold-water coral reefs? Why are there, for example, no more of them in the Norwegian and New Zealand fjords where the distribution of intermediate and deep water masses is right, and where there is ample suitable hard substrate (rock bottom) with high current speeds? However, some recent microbial studies seem to point the way further. Yakimov et al. [42], for example, recently found metabolically active microbial communities associated with deepwater corals in the Mediterranean. Also recent OHI-related research of the microbial food chain surrounding the coral reefs off Mid-Norway has provided some interesting new findings. The contrast between coral-associated and free-living bacteria may suggest that few free-living bacteria are directly ingested by the coral and that instead, corals feed on non-bacterial plankton. Small (100–200 μm) zooplankton has been suggested important in the diet of corals [43]. In addition, the tissue-associated bacterial communities potentially provide a direct translocation of nutrients through metabolism of particulate and dissolved organic matter in the seawater. One *Lophelia pertusa* associate was studied in more detail and named "Candidatus Mycoplasma corallicola" [44]. This bacterium was abundant in *L. pertusa* from both sides of the Atlantic Ocean and is considered an organotrophic commensalist [44]. Given the importance of chemosynthesis in deepwater ecosystem development and functioning, cold-water coral reef communities may be linked to a diversity of chemoautotrophic microorganisms that synthesize organic compounds from inorganic compounds by extracting energy from reduced substances and by the fixation of dissolved CO_2. Just a tiny fraction of microorganisms associated with deepwater coral reefs have yet been identified, and even less assigned to a function. Although no nutritional symbiosis based on chemosynthesis [45] are known to have been documented on deepwater coral reefs, primary producers affiliated with chemoautotrophs (utilizing H_2S, NO_2^-) and methanotrophs (utilizing CH_4) have been found associated with the reef animals and their ambient environment [41, 42, 46]. Thus, also light hydrocarbons can probably stimulate the growth and the high biodiversity found on the *Lophelia* reefs associated with some Norwegian hydrocarbon fields [47]. Only further detailed studies of the reefs will be able to answer these important questions.

The cold-water coral reefs and carbonate mounds represent exceedingly valuable habitats for numerous species, besides the corals themselves. Also for fish, they represent shelter and nursing homes for juveniles, as one of the few comparative studies of on-reef versus off-reef fish-counts for deepwater coral reefs documents. Fish species' richness and abundance was found to be greater on the reef than over the surrounding seabed, as 92% of species, and 80% of individual fish were associated with the reef. The results indicate that the reefs have a very important functional role in deepwater ecosystems as fish habitats [48]. In particular, visual monitoring of coral reefs at the Morvin and Kristin fields, off Mid-Norway, has shown that the redfish (*Sebastes* sp.) probably spawns at the *Lophelia* colonies (Fig. 5). During the month of May, numerous fish with wide bellies, obviously ready to spawn (this species spawns live sprats), congregate very close to the *Lophelia* colonies (Fig. 5). The juveniles can immediately after spawning find full protection against predator fish within the complex structure of the live and dead *Lophelia* skeleton mesh-work. Furthermore, the Norwegian coral reefs have long been known to fishermen as "Uer-bakker" [26], meaning "Redfish slopes," and Furevik et al. [49] were the first to report scientific evidence that long-line catches of redfish, ling, and tusk can be significantly greater on the reefs than in off-reef areas. In addition, Husebø et al. [50] set long-lines in coral habitats and found significantly more fish than



Marine Life Associated with Offshore Drilling, Pipelines, and Platforms. Figure 5
Fish sightings on *Lophelia*-reefs at Norwegian hydrocarbon fields Kristin and Morvin. *Upper and lower left*: *Sebastes* sp. (redfish) congregating for spawning on a 1.5 m tall *Lophelia* colony. Upper right: Seithe feeding near a *Lophelia*-reef at Kristin. The fish often swims around the ROV when operating off Mid-Norway. *Lower right*: A monk fish resting on the seafloor next to a *Lophelia* colony at Morvin. Notice its white protrusions under its head that resemble the *Lophelia* branches. This fish is sometimes found to be resting on top of *Lophelia* colonies, where it is well camouflaged

elsewhere and also that the fish were generally larger than those caught in the non-coral areas. The importance of these habitats and their internal ecological dynamics has been discussed in more detail by Buhl-Mortensen et al. [24].

According to authoritative assessments of possible threats to the cold-water coral reefs, the main threats according to Roberts et al. [51] are: (1) Bottom trawling, (2) Hydrocarbon drilling and seabed mining, and (3) Ocean acidification (i.e., global climate change [52, 53]). In addition, for each individual coral reef, there is always the possibility of dramatic environmental changes by natural causes, such as nearby underwater avalanches (burial), and in the case of the hydraulic theory being viable, that the seepage or venting is naturally depleted or exhausted, and becomes "turned off." However, the OHI seems to come out of such assessments far better than the fishery industry: "Compared with widespread evidence for physical damage to reef structures from bottom trawling, there is little evidence that hydrocarbon exploitation substantially threatens cold-water coral ecosystems. *L. pertusa* colonizes North Sea oil platforms and seems to have a self-seeding population, despite proximity to drilling discharge. Greatest concern is over the potential for drill cuttings to smother reef fauna, but such effects would be highly localized when compared with the extent of seabed affected by bottom trawling" [51]. This latter scenario was carefully tested by a 3-month-long drilling and monitoring campaign of four production wells at the Morvin field off Mid-Norway, where numerous coral reefs exist on-site. The results of this campaign are reviewed in a later chapter.

To sum up the threats, based on OHI-related observations and modern publications, it is found that the coral reefs thrive, despite them being located close to OHI-structures and to sporadic drilling activity. They even colonize parts of the OHI-platforms, as seen in Fig. 5, from the Statfjord field [47, 54]. The main threats to the deepwater coral reefs are therefore mostly mechanical, as they are delicate structures and cannot sustain the mechanical indiscriminate stress imposed by portions of the fishing industry and to a much lesser extent, the OHI. This has been amply documented off Norway and off New Zealand [55–57]. Furthermore, based on abundant visual and geophysical data mainly acquired from the OHI, it was possible for Norwegian authorities to be first to officially protect and conserve a large portion of cold-water coral reefs. The first area to be protected against bottom trawling was the Sula Reef in the late 1990s and the Haltenpipe reefs off Mid-Norway with a 970 sq km large protected area [36]. Today, awareness of the ecological significance of deepwater corals is growing rapidly, as it is known that colonial corals provide important habitats and could play a critical role in the life history of many marine species, including fish of commercial interest [58, 59]. This awareness has lead to a general call for the establishment of marine protected areas (MPAs) and especially to protect the most important cold-water coral habitats [60].

Marine Life Affected by the OHI

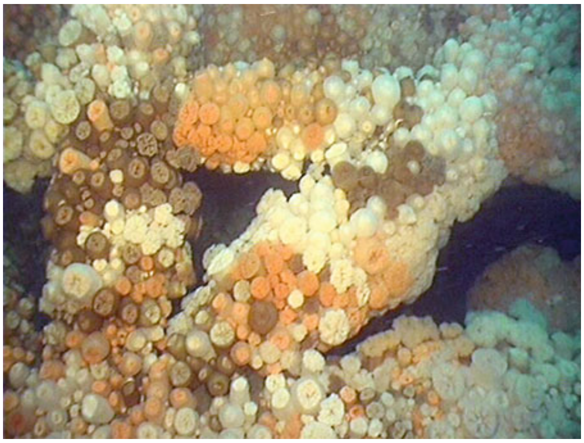
Methods of Observation

Before an offshore hydrocarbon field is developed, rules and regulations say that biological baseline studies shall be carried out. Over the last 2 decades, the focus of such studies has shifted from being purely based on detecting pollution effects from OHI-activity to also include assessments of biodiversity and ecology [61]. A standard outline of a station pattern for sampling seafloor biology is either a grid, covering a large area, or samples obtained along lines forming a cross, with the longest axis downstream with respect to the prevailing current. Several replicate samples are collected at each station. Because this type of sampling was found to be relatively “blind” to any local features on the seafloor, such as pockmarks and cold-water coral reefs, it has now been complemented with an initial seafloor mapping campaign. The macrobenthos fauna of interest in baseline sampling surveys comprises the following main taxonomic groups: Polychaeta, Crustacea, Mollusca, Echinodermata, and Varia (remaining groups). Only animals larger than 1 mm (macrobenthos) are included in such analyses. Subsequent to the development and production at offshore hydrocarbon fields, the sampling of sediments and macrobenthos is repeated perhaps every other year.

The study of benthic communities can provide an indication of pollution from offshore activities. Similarly, the geochemical analysis of sampled sediments can provide levels of pollutants and their distribution around the offshore installations. Any changes in species composition and densities of individuals can also provide tell-tale signals about damaging pollution. Experience from many years of such work has proved that the benthic fauna near OHI-installations can be affected by a number of factors, including the discharges of drilling fluids, cuttings, and others, including the accidental release of oil and also physical disturbances. One of the current challenges of planning future environmental/baseline surveys in the OHI is to improve the cross-departmental communication, as the detailed seafloor mapping is performed by others than those responsible for environmental surveys [61].

During initial mapping and inspection along potential pipeline routes at the Morvin field, off Mid-Norway, numerous coral reefs were found and visually inspected. Here, it was also decided to perform a detailed geochemical investigation prior to the production drilling activity. One of the largest reefs, the MRR (“Morvin Reference Reef”) was systematically investigated and sampled (water, sediment, and organisms). This reef is located inside a large (130 m × 80 m × 10 m) pockmark depression (Figs. 6 and 7). The reef occupies about one third of the pockmark, growing from the maximum depth (at 370 m), up along the northern side, to the rim of the depression (at 360 m water depth). The MRR is about 80 m long, 25 m wide, and spans the elevation interval: 360–370 m. *Lophelia pertusa* is the main reef-building coral [47]. The geochemical analyses of sediments at Morvin proved that they contained varying concentrations of light hydrocarbons (methane–butane). Because the sediments in which these hydrocarbons were sampled are located within the oxygenated upper portion of the seafloor (i.e., only 40 cm below surface), any hydrocarbons remaining there over long time would have been reduced (oxidized) relatively rapidly. Therefore, there must be a natural seepage of light hydrocarbons from below, via molecular and fluid migration. This type of fluid flow is called “micro-seepage” of light

Marine Life Associated with Offshore Drilling, Pipelines, and Platforms. Figure 6
Special features on the seafloor attract and protect large fish. Here is an aircraft wreck from World War II off Mid-Norway, which is now utilized as shelter for large Lings. At least five larger than 1 m length can be seen here. The wreck was detected with side scan sonar and inspected by ROV as part of a site survey for exploration drilling



Marine Life Associated with Offshore Drilling, Pipelines, and Platforms. Figure 7
A subsea steel template structure in the central North Sea, Danish Sector. The steel structure has been completely colonized by sea anemones. This is also a perfect hiding place for fish of all sizes

hydrocarbons [8, 62]. However, because no more than eight samples were acquired at Morvin, it is not possible to state any significant statistical variation over a regionally significant area.

The highest interstitial hydrocarbon concentrations were found in the upper sediments at location S8, which is inside a unit-pockmark, located just to the north of MRR (Fig. 4). Combined with modern results from molecular biological methods, there is here factual support for the notion that a nutrient-rich, "fertile" substratum represents one of the keys to understanding the location of this deepwater coral reef. Previous studies of *Lophelia* O-C stable isotopes [63, 64] show that there are relatively large variations in the δC^{13}-value (ranging from −2‰ to −9‰ PDB). For corals, these negative values have so far been interpreted as being caused by ambient pH-variations, among other factors. However, similar variations in stable isotopes in bivalves at seep sites in the Gulf of Mexico were interpreted as being caused by hydrocarbon uptake in microorganisms, subsequently ingested by the bivalves. This suggests that future studies of the coral reefs should also include systematic stable isotope analyses, combined with more geochemical sediment analyses.

The Impact of Exploratory Drilling

Scientific Drilling When the Deep Sea Drilling Project (DSDP) initiated its research with the drilling vessel "Challenger," in 1974, it found indications of oil at nearly 3 km water depth on its second drilling hole on the first exploration leg (DSDP, Leg 1, Hole 2). When the first cores came onto the deck, there was an unmistakable smell of crude, and small droplets of oil could clearly be seen in the carbonate rock sampled. The objective of this hole was to determine the nature of the so-called Challenger Knoll on the abyssal plain of the mid-Gulf of Mexico. By drilling 144 m into this feature, the scientists determined it was a carbonate capped salt dome. However, because the DSDP was a pure scientific drilling project, it should not be looking for oil. A panel of experts was immediately established to make sure that the coming drilling legs should not target hydrocarbon-prone areas. This is how the "Pollution Prevention and Safety Panel" of the DSDP and later the Ocean Drilling Program (ODP) was borne. Today, this panel of experts of the IODP (Integrated ODP) has been re-named to EPSP (Environmental Protection and Safety Panel).

After exploration drilling for oil had come underway, on the Norwegian Continental Shelf (NCS), during the late 1960s, strict rules and regulations were imposed on the industry. For example, no drilling was allowed on the NCS unless a special "predrilling site survey" was performed. This includes a full seafloor coverage of side scan sonar acoustic images, which is able to detect all types of features on the seafloor larger than about 3 m in extent, such as wrecks, munition, rock outcrops, and many other features, including man-made features on the seafloor.

OHI-Related Drilling On the Norwegian Continental Shelf, it has been normal practice to discharge the sediments produced from drilling ("cuttings") and drill-mud from drilling the first 600–800 m of the well (the "top-hole") directly to the water column, where they are dispersed by currents. The largest particles (about 90% of the cuttings) will then be deposited in a slight mound within about 200 m of the well location and will form a slight mound on the seafloor. Not only do deepwater coral reefs challenge modern science with numerous questions on how they can exist in the middle of a seemingly barren shelf environment [47] – they also present a challenge on how to drill safely without damaging them. At the Morvin field, mentioned earlier, this was a major challenge to field development.

Production Drilling at Morvin, a Special Case The Morvin-field is more densely populated by significant deepwater coral reefs than any other Norwegian offshore field [47]. Whereas pipeline routes and the final location of steel templates could easily be adjusted to avoid the coral reef occurrences, one of the major challenges was how to drill the production wells without harm to the reefs. Because there are significant deepwater coral reefs within 200 m at both the planned Morvin production templates, it was not possible to use the normal practice without incurring damage to some of the reefs. It was therefore decided to seek authorities' permission for discharging top-hole cuttings some distance away from the templates, at locations deemed safe for the coral reefs. The chosen technical solution was to employ a "Cutting Transportation System" (CTS), consisting of two 600 m long flexible tubes installed on the seabed. They were attached to five gravel filled bags ("bigbags") to prevent them from moving on the seafloor and were attached to a manifold placed on the seafloor. This manifold had one exhaust pipe leading from the wellhead where the cuttings were collected and pumped to the CTS. Currents were predicted and numerically modeled, and the optimal discharge locations were found where cuttings would not do harm to the surrounding significant coral reefs. The cuttings were discharged at relatively high speed through recoil dampers placed about 1 m proud of the seafloor at the end of the CTS-hoses.

The cuttings emitting from the recoil damper were found to be heavier and less fine-grained than predicted. This resulted in the discharge plume emitting as a heavy fluid, and spreading along the seafloor as a turbid and heavy cloud. Because this cloud did not spread high up into the water mass most of it accumulated near the end of the CTS hoses without any damage to the corals. Continuous visual monitoring by ROVs of the CTS and the nearest coral reefs was done during the whole top-hole drilling operation at Morvin. This ROV-work also included lifting and repositioning (adjusting) of the recoil dampers, as the heap of cuttings grew to heights of up to 1 m. A total of three sediment traps were also placed on the seafloor in order of documenting the final distribution of any resuspended cuttings. Three of the downstream coral reefs were continuously monitored by using an automatic time-lapse "satellite photo rig" and by ROV-monitoring. This Morvin experience proved that production drilling can also be done near scattered coral reefs without harming them.

Observations Along Pipelines

Prior to the laying of trunk pipelines on the seafloor, detailed visual surveys are performed. The objectives of these "pre-lay" surveys are to make sure the pipeline does not cross any dangerous areas or any features that can hinder the laying. In addition, all features and marine life are also recorded along the pipeline route. Because underwater visibility is restricted, the visually documented corridor is most often no wider than about 10 m. However, the length of the surveyed corridor, which depends on the length of the pipeline to be laid is anything up to several hundred kilometers. During the last quarter century, 1984–2010, several thousands of km long transects were surveyed by Statoil. Thus, trunk gas and oil pipelines have been laid from the Snehvit field in the southern Barents Sea and from numerous fields offshore Mid- and Southern-Norway. Detailed knowledge about the seafloor and marine life has, therefore, been gathered along transects crossing all ranges of water depths of the North Sea, stretching from Norway to the UK, and from Norway to Germany, Belgium, and France (Dunkerque). In this way, the general seafloor has been imaged, meter for meter, across 50 m iceberg ploughmarks in the Barents and Norwegian Seas, to typically 100 m wide, 5 m deep pockmark craters in the northern North Sea, to wide stretches of rugged sandwave fields off Netherlands and Belgium, to the normal, even, and uneventful seabed in all of the seas. The seabed life pattern always seems to be the same: there is apparently very little variation in macrofaunal life except for when there are special features on the seafloor, such as boulders and rock outcrops, besides the special fluid-flow-related features: pockmark craters and other micro-seeps, and the eventual macro-seeps, which are very rare.

The fact that marine life proliferates wherever there are "special features" on the seafloor means that wherever a new man-made structure is installed on the seafloor, there may be a dramatic impact on the visible macrofaunal life pattern (Fig. 6). On all concrete-coated trunk pipelines laid across the pre-surveyed sections, it is clearly noticed how the pipeline structure introduces new opportunities for benthic life. For example, it only took 1 year before thousands of *Nephrops norvegicus* crustaceans were established along kilometer long stretches of the "Statpipe" pipeline, the first 36″ trunk pipeline laid across the 300 m deep Norwegian Trench (North Sea), in 1985. They were clearly colonizing small sections of the seafloor along the outer, curved wall of the pipeline. Their holes were located snugly inside the sediments nearest to the pipeline wall. The clawed animals had their bodies halfway out of their openings, easily visible during the first thorough visual inspection of the pipe, the so-called as-laid survey, conducted about 1 year after the pipeline was installed. Over time, the annual pipeline inspections, which cover long sections of both sides and the top of pipe, every other year, document not only how the trunk pipelines gradually bury themselves into the seabed, but also how they are colonized by a variety of different invertebrates. Unfortunately, this information is seldom used for quantifiable studies by biologists or marine ecologists, although the data should be fully available for such work. The reason is probably, that this type of visual information is unknown and that it perhaps does not belong in the scientific "vocabulary" of traditional marine biology or ecology.

To illustrate some of the important information content in this unique pipeline inspection data set, there are three particularly noticeable occasions, episodes, and occurrences that are worthy of further description: (1) A temporary anoxic condition recorded in the central North Sea, (2) "Drifts" of dead fish (mackerel) along a pipeline section, and (3) Fish protected by OHI-pipelines.

Temporary Anoxic Conditions

During the annual inspection in 2004, of the Europipe 1 trunk gas pipeline from Kårstø, SW Norway, to Northern Germany, a few dead fish and invertebrates were seen lying on the seafloor along the pipeline about 50 km south of the Draupner platform, central North Sea. There was also some white material, believed to be bacterial mats. This section of the pipe is partly buried into the seafloor and is about 40 cm proud of the smooth seafloor surface. Both sides of the pipeline were affected and, there were also dead invertebrates on top of the pipe itself. Although this anomalous observation could easily have been interpreted as something to do with the pipeline, some further investigations and measurements precluded such a conclusion. Anoxia was obviously the cause of this "mass extinction," – but, what had caused it? Along one short section, it was only the southern side of the pipeline that was affected. This annual inspection was conducted during May, when the expected ambient water temperature should be around 6°C. But, actual measurements showed an ambient water temperature of only 1.1°C near the seafloor. Measurements through the water column showed the lower 20 m of seawater only had between 1.1°C and 2.0°C. This very cold, nearly freezing water apparently occurred as a dome-shaped, dynamically stable water mass occupying the general depression in the seafloor, where the pipeline had been laid. During installation of the Europipe 1 pipeline some years earlier, portions of it were laid into the meandering relict "Elbe valley" on the seafloor of the mid- and southern North Sea. This was obviously where the cold water had accumulated. Investigations of recorded water temperatures further southeast in the German Bight showed that very cold water had formed during the winter months, January and February when there was a long cold spell. Because there had been no subsequent storms, the cold, dense water mass had since followed the deepest local portions of the seafloor, and gradually flowed toward the central portion of the North Sea, where it engulfed parts of the Europipe 1 pipeline and where it became stagnant and anoxic. During the subsequent annual pipeline inspection, 2 years later, there was neither sign of dead animals nor any other evidence of this temporary "extinction event." These observations clearly demonstrate that great variations in the natural seafloor environment, even in the relatively busy central North Sea can occur without being noticed on the sea surface.

"Drifts" of Dead Fish Along a Pipeline

The fact that large trunk pipelines laid onto the even, flat, featureless seafloor act as barriers for objects drifting across the seafloor is well known. During the early years of trunk pipeline construction in Norwegian waters (1983–1997), quite a lot of garbage, that is, paper, plastic bags, bottles, cans, etc, together with natural debris, such as kelp and sponges, etc, was noted to accumulate along sections of the newly laid pipes. During the last decennium, however, the amount of garbage has decreased markedly, probably as a result of stricter garbage handling rules in general, and Norway in particular. However, such rules are not always obeyed, as the following example proves. During the annual inspection of a section of the 36″ trunk gas pipeline, Europipe 2, in the Norwegian sector of the North Sea, a huge and elongated pileup of dead fish (mackerel) was visually recorded. It formed a ca. 3 km long and nearly 1 m high "drift" along the western side of the pipeline. The volume of dead fish was estimated to about 10,000 m^3. Because this freshly killed mackerel was obviously dumped illegally by one or several of about seven nearby fishing vessels, the Norwegian Fishery Directory was contacted, and on the following evening, live video footage of the drift was broadcast as a news item by the directory on national TV. Although this live footage, from one of OHI's annual pipe inspections surely made a public impression, it is very rare that such information is used for public purposes.

Fish Protected by OHI-Pipelines

The largest pipeline constructed to date, in the North Sea and Norwegian Sea, is the 1,200 km long 40″ and 44″ Langeled pipeline, transporting gas from the giant Ormen Lange field off Mid-Norway, via Nyhavna in Mid-Norway and the Sleipner field in the central North Sea to Immingham on the east coast of the UK. During the initial as-laid survey, immediately after laying, in 2006, a large school of newly spawned fish was documented along a 50–100 km long section of the pipe, in UK waters. This school was also observed to swim along the pipeline about 3 months later, after the individual fish had grown significantly larger in size. Trawling is routinely practiced along exposed trunk pipelines. The intent is to catch fish that swim along and hide beneath free-spanning sections of the pipelines. Because this large school of fish had obviously been spawned near the newly laid pipeline, and because it utilized the opportunity for refuge along the pipeline, it was suggested by OHI-personnel that measures to protect the school should perhaps be taken by British Authorities. To protect a corridor of, say 3 km width, along a 100–200 km long section of the Langeled pipeline could have been feasible. The following annual inspections could also have visually documented the fate of such a school of fish. However, it was soon realized that suggestions of this kind were premature, in 2006, as it was not expected that anybody would actually understand what the problem may be, or what the suggestion was really about. This demonstrates clearly how little is known and currently understood about the subsea world and how difficult it is to convey ideas and impressions from this world, unless the receivers have similar or comparable experience.

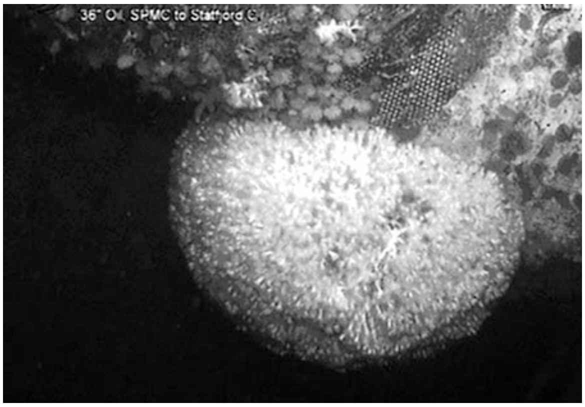
The higher abundance of fish sightings (of fish >0.5 m in length) along trunk pipelines was already noticed after the installation of Statoil's first trunk gas pipeline, the over 250 km long and 36″ Statpipe pipeline from the Statfjord field to Kalstø on the SW coast of Norway. This pipeline crosses the 300 m deep Norwegian Trough, where there is a varying density of pockmark craters in the seafloor. During the very first annual survey of this pipeline, a strong correlation between number of fish sightings and the density of pockmarks was noticed. The fish counted in 1987 were either seen swimming along the sides of the pipeline, or occupying some of the space underneath the pipeline where it had intermittent spans over irregular terrain in the seafloor. This comparison was made with fish sightings in 1984, over the same, undisturbed section of seafloor, before the pipeline was installed. The increased number of fish seen, after the pipeline was installed was dramatic and close to tenfold along some of the seafloor sections. Also before the pipeline was installed, there was evidently many more fish in densely pockmarked areas than in areas without pockmarks. This example shows that trunk pipelines attract fish and also likely protects them to some degree.

The question of protection of trunk pipelines against damage by trawl-board impact was assessed prior to installation of the Statpipe pipeline. Carefully designed laboratory studies in test tanks and rigs, however, soon proved that no serious damage would possibly be inflicted on the concrete coated steel trunk pipelines by normal trawl-boards. Therefore, most such pipelines are left on the seafloor without being actively protected by trenching or gravel cover. In some areas, where the seafloor relief is varied, by numerous pockmarks or iceberg ploughmarks, there may be a need to modify the seafloor topography to avoid long free spans in the pipe. Spans longer than about 80 m are unacceptable for two reasons: (1) motions in the pipe caused by water currents may inflict material damage to the pipe, and (2) trawl-boards may snag underneath the pipe, during fishing along the pipeline. Seafloor preparation and intervention is then performed, either by trenching the seafloor highs or by dumping gravel inside seafloor troughs. This type of intervention is performed before installation of pipeline or immediately after installation. In the German sector of the North Sea, where the water depth is not greater than 150 m, all trunk pipelines have to be buried below the seafloor surface. The annual inspection of buried pipelines is done by the use of induced electric currents and magnetic detection of the buried pipe, combined with visual inspection of the seafloor above the buried pipe. During the last 10–15 years, the annual surveys in these waters have not, unfortunately produced many fish sightings, probably as a consequence of lack of protection and intensive fishing activity.

The Impact of Platforms and Other Fixed OHI-Structures

The six giant concrete platforms Statfjord, A, B, C and Gullfaks A, B, C were installed on the seafloor at the Statfjord and Gullfaks fields in the northern North Sea during a 15 year period in the 1980s and 1990s. Apart from the seafloor at the deepest site, Gullfaks C, the seafloor is dominated by sand, gravel, and patches of boulders. Prior to installation, large fish, such as *Brosme* sp., Ling, Cod, and Seithe, were only occasionally seen swimming around. However, during a detailed preinstallation site survey of the deeper Gullfaks C site, a more complex situation was found. The seafloor at Gullfaks C contrasts with the other sites by having a seafloor covered by soft, fine-grained clay-dominated sediments. Furthermore, there are normal-pockmarks in this area. A dedicated visual survey of pockmark craters here revealed that larger fish, most often the *Brosme* sp., were located inside many of these pockmarks. And inside one particular pockmark, of 8 m depth and 120 m length, there were up to 20 large fish located in the deepest end of the pockmark crater [10]. About fifteen of the large fish (Ling, Cod, and Brosme) were swimming against the southward flowing current, whereas about five fish were found inside a tunnel at the deepest portion of the pockmark [10]. This example clearly demonstrates some of the unexpected heterogeneity of the seafloor and that it takes dedicated visual surveying to unravel the true nature of marine life on the undisturbed seafloor.

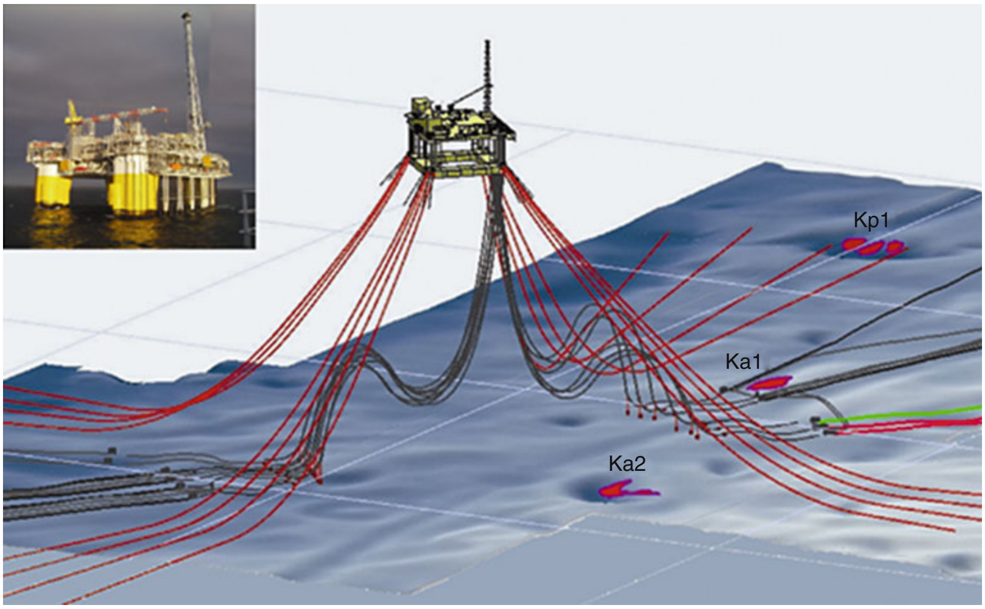
After about 30 years of operating large fixed structures, it is a well-known fact within the OHI that there are plenty of large fish living near the structures. At some of the hydrocarbon fields in the North Sea special trawl-fenders have been installed on the seafloor to prevent trawlers from fishing within the 500 m forbidden safety zone surrounding all fixed platforms which are over the sea surface. These trawl-fenders are robust, up to 3 m high steel poles driven into the seafloor. They are intended to snag the trawl equipment before it can damage any of the production gear and infrastructure placed on the seafloor within the safety zone. It is also a well-known fact that filter-feeders, such as sea anemones and other sessile animals colonize the legs of both steel and concrete platforms (Fig. 7). However, until for about 15 years ago it was not known that colonizing



Marine Life Associated with Offshore Drilling, Pipelines, and Platforms. Figure 8
A relatively small *Lophelia* colony seen on a man-made structure at Statfjord, Norwegian sector of the North Sea. The colony cannot be older than about 22 years (age of structure)

ahermatypic corals also lived on these structures (Fig. 8). Thus, OHI-related fixed structures make ideal artificial reefs on the seafloor.

By law, all the OHI-related structures placed on the seafloor shall be decommissioned after use, that is, they shall be dismantled and disposed of in a safe and environmental friendly manner. However, judging from experiences in the North Sea, even some of the large trunk pipelines could perhaps remain until they either buried themselves or corroded into destruction. In the meantime, they would undoubtedly represent an added biologically friendly asset to the normal seafloor, at least in some places (Fig. 9). For the giant concrete structures, the case is also one of cost benefit. These structures are placed firmly into the ground and will require a lot of energy to remove. Because concrete is a very environmental friendly substance and has a high longevity, it should be seriously debated if they could



Marine Life Associated with Offshore Drilling, Pipelines, and Platforms. Figure 9
A perspective image of how the Kristin semi-floating platform is located above the seafloor. The water depth below the platform is 305 m. *Red lines* indicate anchor chains and wires. The *black lines* indicate flowlines and risers piping products (gas, condensates, etc.). The chains and wires that hold Kristin in place are attached to “suction anchors,” inverted domed steel cylinders sucked into the seafloor clays. The coral reefs, Ka2, Ka1, and Kp1 (in *red*), have been inspected twice before installation and several times after installation of the platform and the other infrastructure on the seafloor. No damage has been found on the reefs. Image is based on true multibeam mapping of the seafloor and digital technical drawings of the platform, shown in a photo *upper left*. This image shows that the platform and infrastructure can be relinquished after use, without any damage to sensitive biology on the seafloor. Courtesy of Statoil ASA and Leslie Austdal

remain in place for 'ever.' They could act as home for plenty of fish, and could even be used by professional fishermen as fish nurseries and fishing grounds.

Future Directions

More dedicated detailed seafloor mapping, combined with visual documentation is undoubtedly needed for improving the understanding and knowledge of life on the seafloor in general. Through the Norwegian "Mareano" project, started by the government in 2005, this type of work has actually started in Norway. Swathe multi-beam detailed seafloor mapping is here combined with up to kilometer long video (visual) transects. In addition to sampling, the seafloor is mapped for biotype and many other parameters. In the long run, such mapping and documentation is necessary for adequate management of the resources in the sea. However, it may take another 100 years for one vessel to map out the whole Norwegian EEZ (Exclusive Economic Zone) in this fashion. This therefore, calls for more technological development and probably also for the use of autonomous robotic systems in the future.

With regard to the OHI-related structures placed on the seafloor, such as trunk pipelines, steel templates, and giant platforms, there needs to be a thorough discussion as to which of them should be relinquished (removed) and which structures could serve to stimulate and protect the marine bio-environment. Judging from the visual documentation over the last 30 years in the North Sea, it can be deemed favorable to let some of the installations remain on the seafloor as artificial reefs and for the future benefit of both marine life and humans.

Bibliography

1. Martin TR, Olsen KR, Cahill MM (2010) Artificial reefs – an important tool for mitigation and restoration. Ocean News Technol 16(4):28–29
2. Maurer BA, McGill BJ (2004) Neutral and non-neutral macroecology. Basic Appl Ecol 5:413–422
3. Humphris SE, Zierenberg RA, Mullineaux LS, Thomsen RE (eds) (1995) Seafloor hydrothermal systems: physical, chemical, biological, and geological interactions. American Geophysical Union, Geophysical monograph 91, 510 pp
4. Etter RJ, Rex MA (1990) Population differentiation decrease with depth in deep-sea gastropods. Deep-Sea Res 37:1251–1261
5. Rex MA (1983) Geographic patterns of species diversity in the deep-sea benthos. In: Rowe GT (ed) The sea. Wiley, New York
6. Piepenburg I et al (2001) In: Schăfer S et al (eds) The northern north Atlantic – a changing environment. Springer, Berlin
7. Schäfer P, Ritzrau W, Schlüter M, Thiede J (2001) The northern north Atlantic – a changing environment. Springer, Berlin, 500 pp
8. Judd AG, Hovland M (2007) Submarine fluid flow, the impact on geology, biology, and the marine environment. Cambridge University Press, Cambridge, 475 pp
9. Suess E (2010) Marine cold seeps. In: Timmis KN (ed) Handbook of hydrocarbon and lipid microbiology, vol 1. Springer, Berlin, pp 187–203 (Part 3)
10. Hovland M, Judd AG (1988) Seabed Pockmarks and Seepages. Impact on geology, biology and the marine environment. Graham & Trotman, London, 293 pp
11. Hovland M, Thomsen E (1989) Hydrocarbon-based communities in the North Sea? Sarsia 74:29–42
12. Hovland M (2002) On the self-sealing nature of marine seeps. Cont Shelf Res 22:2387–2394
13. Wegener G, Shovitri M, Knittel K, Niemann H, Hovland M, Boetius A (2008) Biogeochemical processes and microbial diversity of the ullfaks and Tommeliten methane seeps (northern North Sea). Biogeosciences 5(4):1127–1144
14. Hovland M (2007) Discovery of prolific natural methane seeps at Gullfaks, northern North Sea. Geo-Marine Lett. doi:10.1007/s00367-007-0070-6
15. Seibel BA, Dierssen HM (2009) Animal function at the heart (and gut) of oceanography. Science 323:343–344
16. Moore CJ (1999) Seeps give a peek into plumbing, explorer. Am Assoc Petrol Geol 99:22–23
17. Dimitrov LL (2002) Mud volcanoes – the most important pathway for degassing deeply buried sediments. Earth Sci Rev 59:49–76
18. Dupré S, Woodside J, Klaucke I, Mascle J, Foucher J-P (2010) Widespread active seepage on the Nile Deep Sea Fan (offshore Egypt) revealed by high-definition geophysical imagery. Mar Geol 275:1–19
19. Vogt PR, Crane K, Pfirman S, Sundvor E, Cherkis N, Flemming H, Nishimura C, Shor A (1991) SeaMarc II sidescan sonar imagery and swath bathymetry in the Nordic basin. EOS Trans 72:486
20. Hovland M, Hill A, Stokes D (1997) The structure and geomorphology of the Dashgil mud volcano. Azerbaijan Geomorphol 21:1–15
21. King LH, MacLean B (1970) Pockmarks on the Scotian shelf. Geol Soc Am Bull 81:3142–3148
22. Cathles LM, Su Z, Chen D (2010) The physics of gas chimney and pockmark formation, with implications for assessment of seafloor hazards and gas sequestration. Mar Petrol Geol 27:82–91
23. Hovland M, Heggland R, de Vries MH, Tjelta TI (2010) Unit pockmarks and their potential significance for prediction of fluid flow. J Mar Petrol Geol 27:1190–1199

24. Buhl-Mortensen L, Vanreusel A, Gooday AJ, Levin LA, Priede IG, Buhl-Mortensen P, Gheerardyn H, King NJ, Raes M (2010) Biological structures as a source of habit heterogeneity and biodiversity on the deep ocean margins. Mar Ecol 31: 21–50
25. Gunnerus JE (1768) Om nogle Norske coraller. In: Kongelige Norske Videnskabers Selskabs Skrifter, vol 4, pp 38–73
26. Dons C (1944) Norges korallrev. Norsk Vidensk Selsk Trondheim Forh 16A:37–82
27. Wilson JB (1979) The distribution of the coral Lophelia pertusa (L.) [L. Prolifera (Pallas)] in the north-east Atlantic. J Mar Biol Assoc UK 59:149–164
28. Wilson JB (1979) "Patch" development of the deep-water coral *Lophelia pertusa* (L.) on rockall bank. J Mar Biol Assoc UK 59:165–177
29. Hecker B, Blechschmidt G, Gibson P (1980) Epifaunal zonation and community structure in three Mid- and North Atlantic Canyons. Contract report BLM AA551-CT8-49 prepared by Lamont-Doherty for US Department of Interior
30. Zibrowius H (1980) Les Scléractiniaires de la Méditerranée et de l'Atlantique nord-oriental. Memoires de l'Institute Oceanographique 11:247
31. Hovland M (1990) Do carbonate reefs form due to fluid seepage? Terra Nova 2:8–18
32. Hovland M, Croker PF (1993) Fault-associated seabed mounds in the Porcupine Basin, offshore Ireland. Expanded abstract. In: Proceedings of the 55th EAEG Ann. Mtg., Stavanger, Norway
33. Hovland M, Croker P, Martin M (1994) Fault-associated seabed mounds (carbonate knolls?) off western Ireland and North-West Australia. Mar Petrol Geol 11:232–246
34. Freiwald A, Henrich R, Pätzold J (1997) Anatomy of a deep-water coral reef mound from Stjernsund, west Finnmark, northern Norway. In: James NP, Clarke JAD (eds) Cool-water carbonates. Soc Sediment Geol (SEPM), Special Publ 56:140–161
35. Henriet J-P, De Mol B, Pillen S, Vanneste M, Van Rooij D, Versteeg W, Croker PF, Shannon PM, Unninthan V, Bouriak S, Chachkine P (1998) Gas hydrate crystals may help build reefs. Nature 391:648–649 (Porcupine-Belgica Shipboard Party)
36. Hovland M, Mortensen PB (1999) Norske korallrev og prosesser I havbunnen (Norwegian coral reefs and seabed processes), John Grieg, Bergen, Norway, 167 pp (in Norwegian with English summary)
37. Armstrong CW, van der Hove S (2007) The formation of policy for protection of cold-water coral off the coast of Norway. Internal Report, University of Tromsø
38. Fosså JH, Mortensen PB (1998) Artsmangfoldet på *Lophelia*-korallrev og metoder for kartlegging og overvåkning. The biodiversity on Lophelia-reefs and methods for mapping and monitoring. Fisken og Havet 17:1–95 (in Norwegian)
39. Willison JHM, Hall J, Gass SE, Kenchington ELR, Butler M, Doherty P (eds) (2001) In: Proceedings of the first international symposium on deep-sea corals, Ecology Action Centre and Nova Scotia Museum. Halifax, Canada
40. Hovland M, Risk M (2003) Do Norwegian deep-water coral reefs rely on seeping fluids? Mar Geol 198:83–96
41. Jensen S, Neufeld JD, Birkeland N-K, Hovland M, Murrell JC (2008) Insight into the microbial community structure of a deepwater coral reef environment. Deep-Sea Res I 55:1554–1563
42. Yakimov MM, Cappello S, Crisafi E, Tursi A, Savini A, Corselli C, Scarfi S, Giuliano L (2006) Phylogenic survey of metabolically active microbial communities associated with the deep-sea coral *Lophelia pertusa* from the Apulian plateau, Central Mediterranean Sea. Deep-Sea Res I 53:62–75
43. Sorokin YuI, Sorokin Yu P (2009) Analysis of plankton in the southern Great Barrier Reef: abundance and roles in throphodynamics. J Mar Biol Assoc UK 89:235–241
44. Neulinger SC, Gärtner A, Järnegren J, Ludvigsen M, Lochte K, Dullo W-C (2008) Tissue-associated "*Candidatus Mycoplasma corallicola*" and filamentous bacteria on the cold-water coral *Lophelia pertusa* (Schleractinia). Appl Environ Microbiol 75:1437–1444
45. Tavormina PL, Ussler W, Orphan VJ (2008) Planktonic and sediment-associated aerobic methanotrophs in two seep systems along the North American margin. Appl Environ Microbiol 74:3985–3995
46. Penn K, Wu D, Eisen JA, Ward N (2006) Characteristics of bacterial communities associated with deep-sea corals on Gulf of Alaska seamounts. Appl Environ Microbiol 72:1680–1683
47. Hovland M (2008) Deep-water coral reefs – unique biodiversity hot-spots. Springer Praxis, Chichester, 278 pp
48. Costello MJ, McRea M, Freiwald A, Lundälv T, Jonsson L, Bett BJ, van Weering TCE, de Haas H, Roberts MJ, Allen D (2005) Role of cold-water coral *Lophelia pertusa* coral reefs as fish habitat in the North East Atlantic. In: Freiwald A, Roberts M (eds) Cold-water corals and ecosystems. Springer, Heidelberg, pp 771–805
49. Furevik D, Nøttestad L, Fosså JH, Husebø A, Jørgensen S (1999) Fiskefordeling i og utenfor korallområder på Søregga. Fisken og Havet no 15, 33 pp
50. Husebø A, Nøttestad L, Fosså JH, Furevik D, Jørgensen SB (2002) Distribution and abundance of fish in deep-sea coral habitats. Hydrobiologia 471:91–99
51. Roberts JM, Wheeler AJ, Freiwald A (2006) Reefs of the deep: the biology and geology of cold-water coral ecosystems. Science 312:543–547
52. Turley C, Blackford J, Widdicombe S, Lowe D, Nightingale PD, Rees AP (2006) Reviewing the impact of increased atmospheric CO_2 on oceanic pH and the marine ecosystem. In: Schnellnhuber HJ, Cramer W, Nakicenovic N, Wigley T, Yohe G (eds) Avoiding dangerous climate change. Cambridge University Press, Cambridge, pp 65–70
53. Turley CM, Roberts JM, Guinotte JM (2007) Corals in deep-water: will the unseen hand of ocean acidification destroy cold-water ecosystems? Coral Reefs. doi:10.1007/s00338-007-0247-5

54. Gass SE, Roberts JM (2006) The occurrence of the cold-water Lophelia pertusa (Scleractinian) on oil and gas platforms in the North Sea: colony growth, recruitment and environmental controls on distribution. Mar Pollut Bull 52:549–559
55. Roberts JM, Long D, Wilson JB, Mortensen PB, Gage JD (2003) The cold-water coral *Lophelia pertusa* (Scleractinia) and enigmatic seabed mounds along the north-east Atlantic margin: are they related? Mar Pollut Bull 46:7–20
56. Koslow JA, Gowlett-Holmes K, Lowry JK, O'Hara T, Poore GCB, Willimams A (2001) Seamount benthic microfauna off southern Tasmania: community structure and impacts of trawling. Mar Ecol Prog Ser 213:111–125
57. Fosså JH, Mortensen PB, Furevik DM (2000) *Lophelia*-korallrev langs norskekysten. Forekomst og tilstand. Lophelia coral reefs along the Norwegian coast. Occurrence and conditions. Fisken og havet, 2, 94 pp (in Norwegian)
58. Rogers AD (1999) The biology of *Lophelia pertusa* (Linnaeus 1758) and other deep-water reef-forming corals and impacts from human activities. Int Rev Hydrobiol 84:315–406
59. Mortensen PB, Fosså JH (2006) Species dieversity and spatial distribution of invertebrates on *Lophelia* reefs in Norway. In: Proceedings of the 10th international coral reef symposium. Okinawa, Japan, pp 1849–1868
60. Hall-Spencer J, Allain V, Fosså JH (2002) Trawling damage to Northeast Atlantic ancient coral reefs. Proc Royal Soc Lon Ser B 269:507–511
61. Myhrvold A, Hovland M, Nøland S-A (2004) Baseline and environmental monitoring in deep water – a new approach. In: Seventh international SPE conference on health, safety, and environment, Calgary, 29–31 Mar 2004. Paper no. SPE 86776
62. Etiope G, Feyzullayev A, Baciu CL (2009) Terrestrial methane seeps and mud volcanoes: a global perspective of gas origin. Mar Petrol Geol 26:333–344
63. Mikkelsen N, Erlenkauser H, Killingley JS, Berger WH (1982) Norwegian corals: radiocarbon and stable isotopes in Lophelia pertusa. Boreas 5:163–171
64. Mortensen PB, Rapp HT (1998) Oxygen- and carbon isotope ratios related to growth line patterns in skeletons of *Lophelia pertusa* (L) (Anthozoa: Scleractinia): implications for determination of linear extention rates. Sarsia 83: 433–446

Web Sites

www.mareano.no
www.npd.no
www.nerc.ac.uk
www.iodp.org
www.deepseadrilling.org
www.offshoremagazine.com
www.serpentproject.com
www.diverdiscover.whoi.edu
www.statoil.com
www.eu-hermes.net

Marker-Assisted Breeding in Crops

ROBERTO TUBEROSA
Department of Agroenvironmental Science and Technology, University of Bologna, Bologna, Italy

Article Outline

Glossary

Backcross Procedure used by plant breeders to introgress an allele at a locus of interest (e.g., disease resistance) from a donor parent to a recurrent parent, usually a successful cultivar. The recurrent parent is crossed several times to the original cross and selection is performed at each cycle to recover the plants with the desired allele and the largest portion of the genome of the recurrent parent.

Candidate gene A coding sequence that is supposed to be causally related to the trait under selection. The candidate-gene approach is best applied with simple biochemical traits when a clear cause-effect relationship can be established between the gene function and the target trait.

Epistasis The interaction between two or more genes to control a single phenotype. Interaction between two or more loci that control the same trait. The presence of epistatic loci makes it more difficult to predict the phenotypic value of progeny derived either from crosses or from selfing.

Forward genetics Approaches to dissect the genetic makeup of traits starting from the observation of the phenotype. QTL mapping and positional cloning are examples of forward genetics to investigate quantitative traits.

Haplotype Chromosome fragment of varying length carrying a common set of marker alleles in close linkage at adjacent loci. When using haplotypes in association mapping studies, the information of several linked bi-allelic markers is combined as a single, multi-locus informative marker.

Heritability The portion (from 0% to 100%) of phenotypic variability that is genetically determined. The additive portion (i.e., not due to dominance) of variability is inherited from one generation to the next and is the main determinant of the gain from selection. Heritability is specific to a particular population in a particular environment.

Introgression library lines (ILLs) A collection of lines (ca. 80–100) obtained by subsequent backcrosses of a recurrent parent (usually an elite cultivar) with a donor parent, usually a line highly diversified from the recurrent parent for one or more traits. Each ILL carries a fragment (from ca. 20 to 40 cM) of the donor genome different from that carried by the other lines. Collectively, the fragments of all ILLs cover the entire genome with partial overlap. ILLs are ideal for the fine mapping and cloning of major loci and to investigate epistatic interactions.

Linkage disequilibrium (LD) The level of nonrandom assortment of alleles at different loci. The level of LD varies greatly according to the species and the mode of reproduction.

Linkage drag The negative phenotypic effects (e.g., lower yield) on the recurrent parent associated with the loci of the donor parent tightly linked to the locus of interest being backcrossed.

Logarithm of the odds ratio (LOD) A logarithmic value (base 10) of the ratio between the probability of the presence of a QTL vs. its absence. A LOD value of 3.0 indicates that the probability of the presence of the QTL is 1,000-fold higher than its absence.

Metanalysis A comprehensive analysis based on the data of several mapping populations of the same species. The objective is to obtain a better resolution of the LOD profile of the QTLs for the traits of interest.

Near isogenic lines (NILs) A set of two or more inbred lines that share most of the genome except for a small portion that contains functionally different alleles at the target locus. NILs are commonly used for the positional cloning of a locus of interest.

Phenotypic selection Selection based on the observation of the phenotype at different levels of functional organization based on the target trait(s). If the selected trait is highly influenced by environmental conditions and has low heritability, the effectiveness of phenotypic selection quickly decreases.

Pleiotropy Condition where a single locus controls more than one trait. It is more common for biochemical traits.

Positional cloning A series of procedures to clone a locus of interest. Positional cloning is based on the joint analysis of phenotypic data and genotyping profiles of near isogenic material with recombination events at the target region.

Quantitative trait locus A portion of DNA that influences the expression of a quantitative trait. The presence of QTLs is determined through appropriate statistical analysis of phenotypic and molecular data of a mapping population (e.g., linkage mapping) or a collection of unrelated genotypes (e.g., association mapping).

Recombinant inbred lines A collection of homozygous lines (usually from 150 up to 400) obtained following subsequent selfings (usually four or five) of an equivalent number of randomly chosen F_2 plants.

Reverse genetics An approach for discovering the function of a locus by analyzing the phenotypic effects of specific sequences obtained by DNA sequencing. Reverse genetics attempts to connect a given genetic sequence with specific effects on the organism.

Synteny The physical colocalization of linked loci on the same chromosome among different species. Study of synteny can show how the genome of phylogenetically related species has evolved from a common ancestor (e.g., rice for cereals) through rearrangements of the genome (e.g., translocations, inversions, duplications, etc.) in the course of evolution.

Definition of the Subject

Attaining global food security by means of increased crop productivity will require an increase in gains from selection achieved through conventional breeding. To this end, the identification of molecular markers associated with loci controlling traits of agronomic interest coupled with the exploitation of marker-assisted breeding (MAB) approaches provides the opportunity to accelerate gain from selection. In particular, marker-assisted selection (MAS) and marker-assisted backcrossing have been widely adopted to improve resistance to diseases and other relatively simple traits. Notwithstanding these remarkable achievements, the improvement of yield and other complex quantitative traits via MAB has been marginal, mainly due to the difficulty in identifying major quantitative trait loci (QTLs) with an adequately stable effect across environments and genetic backgrounds. Additionally, the effect of most QTLs affecting yield is too small to be detected with either biparental mapping or association mapping. Genomic selection (GS) circumvents this problem by using an index for the selection of unmapped QTLs of small individual effects but with otherwise sizable effect at the whole plant level when selected together. GS is already having a positive impact on the improvement of crop yield, mainly in the private sector where high-throughput infrastructures allow breeders to handle the large number of molecular datapoints that are required for effectively deploying GS. Ultimately, an effective exploitation of MAB to enhance crop performance will rely on a closer integration between molecular approaches and conventional breeding.

Introduction: Global Food Security and Plant Genomics

During the past century, plant breeders have been very successful in constantly raising crop yields to a level sufficient to meet the global demand in food, feed, and fiber. For wheat and rice, the two most important staples of humankind, the so-called Green Revolution spearheaded by Norman Borlaug, awarded the Nobel Peace Prize in 1970, provides the most spectacular example of the contribution of science toward an improved food security [1, 2]. Similar progress has been achieved also in maize, particularly following the introduction of hybrids [3]. This notwithstanding, during the past decade, the rate of increase in yield in cereals, especially wheat and rice, has not met the global demand [4] as shown by the substantial decrease in the amount of global cereal reserves. Additionally, during the past two decades the number of chronically hungry people has increased and is fast approaching one billion. A number of reasons have contributed to this worrisome scenario that has already sparked food riots (e.g., during the 2007–2008 food crisis and also in 2009) and social unrest in a number of less-developed countries. An even bleaker picture looms on the horizon, when mankind will reach a projected nine billion in 2050. Consequently, an acceleration in the rate of gain in crop yields is a must in order to keep up with the need of a burgeoning population that increasingly seeks a protein-enriched, nutritionally balanced diet. The challenge faced by modern breeders is even more daunting in view of (1) global warming and the consequent increased frequency of drought, floods, high temperatures, etc., (2) the decreased availability of natural resources (e.g., water, fertilizers, arable land, etc.), (3) the increasing cost of fuels, (4) the necessity to safeguard the remaining biodiversity, and (5) the increased societal awareness of the critical need to improve the long-term sustainability of agricultural practices and decrease its impact on the environment. More simply, agriculture will need to produce more with fewer resources and more sustainably.

In this daunting scenario, genomics has ushered in a new breeding paradigm based on molecular approaches and platforms that in some cases have already contributed to accelerate the yield gain commonly achieved through conventional breeding practices [5–13]. However, a more widespread adoption of genomics-assisted selection will require the definition of new strategies based on a more effective integration of conventional and nonconventional breeding approaches as well as agronomic practices [14]. Clearly, a better knowledge of the genetic factors that determine yield and its variability from season to season will be instrumental in devising effective marker-assisted breeding (MAB) strategies for enhancing crop performance under a broad range of environmental conditions. As compared to conventional breeding approaches, MAB approaches offer unprecedented opportunities to dissect the genetic control of traits, particularly those that are quantitatively inherited, such as biomass production, yield, and many other agronomic traits selected by breeders.

Molecular Dissection of the Genetic Control of Traits Governing Crop Performance

The first step for the dissection of the genetic control of traits that govern crop performance is the assembly of a linkage (genetic) map based upon the data of the molecular profiles of the marker loci – from as few as 100 up to several thousand – surveyed in a mapping population, usually comprised of ca. 150–200 genotypes such as F_2 plants, F_3 families, recombinant inbred lines (RILs), doubled haploids (DHs), etc., usually derived from the cross of two parental lines differing for the trait(s) of interest. The assembly of a genetic map is based on the level of linkage disequilibrium (LD, i.e., the level of nonrandom assortment of alleles at different loci) among adjacent marker loci on the same chromosome. Accordingly, mapping the loci that control the target trait is also based on the LD between the locus and nearby markers.

The estimated genetic distance between loci (markers or genes) is a function of the average number of recombination events (i.e., crossing-overs) between them at meiosis. The measuring unit used for expressing the distances among loci along a genetic map is the centimorgan (cM), which defines the interval along which one recombination event is expected to occur per 100 gametes produced at each meiotic cycle (i.e., at each sexually reproduced generation). Because a density of one marker per ca. 10–15 cM is usually sufficient to detect the presence of a functionally polymorphic locus with a major effect on the phenotypic variability of a mapping population, the number of well-spaced markers required to adequately sample the targeted species varies from as little as 100–120 as in the case of rice – one of the crops with the smallest genome size (ca. 0.45 billion bp) – to well over 300 for large genomes such as in bread wheat (ca. 16 billion bp). The desired level of genetic resolution will depend on the objective being pursued and the type of genetic materials being used.

For breeding purposes, a density of one marker every 5–10 cM is sufficient for most applications when dealing with elite cultivars. Nonetheless, for the introgression of a particular gene (e.g., a locus for disease resistance) from a wild relative of the crop to the crop itself, a high resolution is desirable in order to avoid the negative effects of the so-called linkage drag caused by negative effects of wild alleles at the loci closely linked to the one being targeted for introgression. A much higher genetic resolution is required when the goal is the cloning of the sequence that affects the target trait. In this case, the screening of several thousands of individuals is required to reach the desired level of resolution.

Cloning the loci that govern a particular trait can be achieved via either forward- or reverse-genetics approaches, or their combination. While forward genetics focuses on the phenotype as starting point, reverse-genetics approaches rely on sequence and functional information of candidate sequences (e.g., expressed sequence tags: ESTs) that are postulated to play a role in the expression of the target trait [15]. Although most results in the dissection of the genetic basis of crop performance and agronomic traits have been obtained via forward genetics, the use of reverse-genetics approaches in *Arabidopsis* and other model species (e.g., resurrection plants, rice, *Brachypodium*, etc.) has been instrumental to elucidate the genetic networks of the signaling pathways that regulate the adaptive response of plants to abiotic and biotic constraints [16–18]. Notably, the spectacular decrease in sequencing costs [19] and the increased availability of sequence information in public databases make the reverse-genetics approach increasingly attractive and feasible.

Following the assembly of the first genetic maps based on the molecular profiling of RFLPs (restriction fragment length polymorphisms; [20, 21]), the introduction of AFLPs (amplified fragment length polymorphisms; [22]), SSRs (simple sequence repeats; [23]), and DArT (diversity array technology; [24]) markers improved substantially the assembly of genetic maps. More recently, high-throughput platforms based on SNPs (single nucleotide polymorphisms), the most frequent polymorphism in living organisms, have enabled a quantum leap in saturating maps with thousands of markers [25–29]. Notably, the spectacular advances obtained with next-generation sequencing (NGS) technology will soon allow for the resequencing of entire mapping populations and association mapping panels of species for which a template sequence is available, thus providing an almost endless supply of markers [30–34].

Once all the molecular and phenotypic data are available, statistical tests will be applied to verify whether the means of the trait values of the genotypes carrying different alleles at a particular marker are significantly different. A test statistic larger than a threshold value rejects the "null hypothesis" (i.e., the mean is independent of the genotype at a specific marker locus) and implies a significant association between the investigated marker and a linked locus that affects the phenotypic value of the target trait. The exploitation of syntenic relationships among phylogenetically related crops has greatly contributed to the identification of additional markers at target regions [35–37] and, most importantly, candidates for the investigated traits, particularly when the genome sequence of one or more of the syntenic species becomes available. This is the case of cereals, where the annotated sequence for rice, *Brachypodium*, sorghum, and maize has allowed for the identification of conserved orthologous set (COS) markers from ESTs that have maintained their microlinearity throughout evolution and speciation [37]. These markers are particularly valuable to assess the possible role of candidate genes in species not yet sequenced (e.g., wheat) and to identify orthologous sequences that have maintained their functions and colinearity across species. Thus, a good understanding of the syntenic relationships at regions underlining a QTL for rather simple traits can provide excellent clues to pinpoint the most likely candidate.

Notably, mapping loci controlling the target traits allows breeders to implement marker-assisted selection (MAS) on the basis of the polymorphic molecular markers flanking the relevant loci. Traits are usually categorized as monogenic (qualitative or Mendelian traits controlled by a single locus) and polygenic (or quantitative; controlled by many loci), the latter being highly influenced by environmental conditions and considerably more difficult to improve consequent to their lower heritability, [38]. Quantitative traits (e.g., flowering time, plant height, biomass production, yield, etc.) are particularly important for breeding purposes. Although the genetic dissection of both qualitative and quantitative traits relies on similar principles, the latter requires more extensive phenotyping and much larger mapping populations.

The prevailing assumption in the field of quantitative genetics has been that continuous variation in trait performance is caused by the segregation and action of multiple genes with a rather similar effect on the phenotype, together with a major influence of the environment which acts like some sort of "statistical fog" that blurs and limits our capacity to identify the genes that control the target trait. These genes, also referred to as polygenes, are known as quantitative trait loci (QTLs; [39]). Although the original concept – but not the acronym – of QTL mapping was first suggested in 1923 [40], the dissection of quantitative traits became eventually possible in the 1980s and the 1990s with the introduction of molecular marker platforms that allowed for genome profiling with the needed level of genetic resolution [41–45]. Two decades of dedicated experiments indicate that most QTL effects are of small magnitude as originally predicted by the so-called infinitesimal model [38, 46, 47]. This notwithstanding, a limited number of so-called major QTLs have shown a rather large effect and, in a number of cases, have been cloned [48, 49]. Once a QTL has been cloned, both genomics and genetic engineering offer additional opportunities for tailoring improved cultivars and crossing reproductive barriers among species, thus expanding the repertoire of genes available to breeders. In view of the importance of quantitative traits in breeding activities and crop performance, particular attention should be devoted to QTL mapping and the implementation of MAB for this category of traits.

Biparental Linkage Mapping

The early studies in QTL mapping were conducted based on the analysis of the means at single markers using simple test statistics, such as linear regression, *t*-test, and analysis of variance. Because a genome-wide survey typically involves a large number of markers, the probability of detecting one or more false positives at the whole-genome level quickly increases unless the threshold of significance is adequately readjusted according to the number of tested markers [50]. Typically, a threshold level of $P_{0.05}$ entails a false-positive discovery rate (i.e., declaring the presence of a locus able to affect the target trait when actually there is no locus) of approximately 5%. Consequently, a mapping experiment based on 100 markers tested at $P_{0.05}$ will identify, on average, five markers putatively associated with loci even when no real locus segregates in the population. In order to avoid this problem, the significance threshold is corrected accordingly through a multiple test adjustment (e.g., Bonferroni's or Tukey's) that will adjust the *P* level according to the number of independent statistical tests that are performed. This notwithstanding, a much more critical shortcoming of this single-marker approach is that no information is provided on the most likely position of the locus and its effects on the phenotype. Due to these major limitations, single-marker analysis was quickly replaced by interval mapping and similar methods based on the estimated linear order of markers on a genetic map. In comparison to single-marker analysis, interval mapping provides a much more accurate estimate of the position and genetic effects of each locus [51–53]. In interval mapping, statistical methods are applied to test for the likelihood of the presence of a QTL. The result of the likelihood tests carried out at regular intervals across the ordered markers is expressed as LOD (Logarithm of the ODds ratio) scores, computed as the $\log_{10}$ of the ratio between the chance of a real QTL being present given the phenotypic effect measured at that position, divided by the chance of having a similar effect when no QTL is present. Thus, LOD values of 2.0 and 3.0 indicate that the presence of the QTL is 100- and 1,000-fold more likely than its absence, respectively. The graphical output is an LOD profile that allows one to compute an empirical confidence interval (usually computed as LOD – 1) around the QTL peak. In order to avoid declaring false-positive QTLs (i.e., declaring the presence of a QTL when the QTL is actually absent), a reasonably high threshold value for the LOD score should be set (usually > 2.5). Iterative software based upon resampling procedures provides a more accurate estimate of threshold values according to the size of the mapping population and the number of markers [54].

Epistasis can greatly influence the outcome of interval mapping. This problem can be partially overcome with the use of composite interval mapping, a statistical procedure that can account for the effects of other QTLs inherited independently from the interval (i.e., chromosome region) being considered, thus reducing the possibility of detecting "ghost" (i.e., false) QTLs. Compared to single-QTL interval mapping, statistical approaches for locating multiple QTLs are more powerful because they can differentiate between linked and/or interacting QTLs that will otherwise go undetected when using single QTL interval mapping. Given the potential impact of epistasis on the response to selection, quantifying its influence on target traits is an important component for designing and organizing any MAS strategy [55]. It is likely that the incorporation of epistatic interactions into more properly devised statistical models will play a relevant role in explaining complex regulatory networks governing the expression of quantitative traits.

A major shortcoming of QTL studies is the low accuracy in detecting the real number of QTLs affecting the genetic variation of the investigated traits, particularly with populations of less than 150–200 families, which is the case in the majority of QTL studies reported so far. A simulation study applied to experimental data showed that with populations of ca. 100–200 families only a modest fraction of QTLs was identified; furthermore, the effect of each single QTL was usually overestimated [56]. Another study showed that detection of QTLs of small effect is very difficult with mapping populations with less than 500 families [44]. These predictions were supported in experiments carried out with maize mapping populations large enough (>400 families) to allow for a meaningful subsampling [57, 58]. Therefore, the chance of detecting a QTL in several environments is small even in the absence of QTL × Environment (QTL × E) interaction. Accordingly, inconsistency of QTL detection across environments has been repeatedly reported [59, 60].

Association Mapping

In the past decade, as an alternative to linkage mapping with biparental populations, association mapping based on the evaluation of panels of unrelated accessions (ca. 150 or more) has been adopted as an additional option for trait dissection [61–65]. The assumption underlying the use of association mapping to detect the presence of loci influencing the target trait is that alleles at two closely linked loci share a historical ancestor, and this original co-occurrence will gradually decay in the population due to recombination events during subsequent meioses. Consequently, the relative allele distributions of an unknown gene and that of a closely linked marker will be nonrandom because the two are in LD. A major factor to be considered for a correct application of association mapping is the presence of population structure, which will significantly bias the results and inflate spurious marker-trait associations (i.e., declaring false positives). Algorithms and methods are being developed to correct for these effects. An important advantage of association mapping is that the linkage is evaluated over the large number of historic meiosis, which in turn entails a much lower LD and higher genetic resolution as compared to linkage mapping with biparental populations. Another advantage is that the genetic variability explored by a large panel of unrelated accessions is much larger than that present in a segregating population derived from two parental lines. Conversely, a major shortcoming of association mapping is that it does not allow for the detection of the effect that a rare, but otherwise agronomically valuable, allele may have on the target trait. In fact, the statistical procedures used for revealing the effects associated to a particular locus/haplotype consider only alleles with a frequency higher than 10% over the entire population; alleles with a frequency lower than 10% are considered rare and as such, are discarded. The cutoff threshold of 10% has been introduced to reduce the ascertainment bias that a small sample (i.e., less than 10%) of accessions would inevitably introduce, being unable to correctly represent the effect of that particular allele at the level of the entire population [62]. Clearly, this is not an issue when dealing with mapping populations where allelic frequencies are expected to be equal to ca. 50%, barring the presence of genetic factors that might influence the transmission of gametes carrying the different parental alleles. In association mapping, the procedure of discarding the individuals carrying rare alleles inevitably reduces the statistical power to identify the role of such loci in controlling the variability measured for the target trait. An example of this has recently been reported in durum wheat, where a locus with a large effect on yield in a biparental cross [162] showed no appreciable effect in a parallel association mapping study where only one of the parental alleles was considered, due to the fact that the other parental allele was present in low frequency [65].

The main factors to be carefully considered for optimizing the effectiveness of association mapping are the level of LD among the investigated accessions and the presence of population structure that could greatly increase the false-discovery rate (i.e., type-I error). Closely related to the concept of LD is the concept of "haplotype," which can be defined as the chromosome fragment carrying a common set of marker alleles in close linkage at adjacent loci [66]. When using haplotypes in association studies, the information of several linked bi-allelic markers is combined as a single, multi-locus informative marker. Haplotypes can be generated *in silico* from sequences deposited in the database, by resequencing target loci (sequence haplotypes) or genetic maps (marker haplotypes). Therefore, haplotypes will extend according to the level of LD, the value of which varies greatly (up to 100-fold or even more) not only among species, but also within a single species according to the frequency of crossing-over events in each chromosome region. As an example, centromeric regions are characterized by very low recombination if compared to subtelomeric, gene-rich regions. Populations characterized by high LD (i.e., extending for > 1 cM, corresponding to several million base pairs (bp) depending on the ratio of the genetic and physical distance) are best suited for a genome-wide search [65]. Alternatively, the utilization of panels with a low LD (i.e., extending < 10,000 bp, typically a small fraction of 1 cM), a condition that is typical of allogamous species like maize [67], allows for a much higher level of genetic resolution and for the validation of a candidate sequence. Clearly, the level of LD influences the number of markers/cM required to obtain meaningful information. As compared to a low LD condition, a high LD level is associated with a proportionally longer haplotype, hence requiring a lower number of markers to conduct meaningful genome-wide surveys. This feature is more prominent in elite materials that have undergone high selection pressure as a result of modern breeding practices, which in most cases has led to a reduction of haplotype diversity as compared to locally grown landraces and, more notably, wild relatives of crops that have not gone through the domestication bottleneck. As an example, LD in wheat – a selfing species that has undergone a very stringent selection mostly due to the importance of quality parameters required by the food industry – extends up to 5–10 cM [65], while in outcrossing species like maize LD is usually below a fraction of cM or even less than 10,000 bp [68]. An example of the high level of genetic resolution made possible through association mapping is shown by the fine mapping and, in one case, positional cloning of QTLs for flowering time in maize [67, 68]. In particular, association mapping revealed that the most important QTL for flowering time per se (i.e., independently from photoperiod sensitivity) in maize corresponds to a 2.3 kb, noncoding, long-distance enhancer region located 70 kb upstream of a gene known to regulate flowering time also in *Arabidopsis* [49]. Another remarkable example in which the functional polymorphism responsible for phenotypic variability was assigned to a noncoding region far (ca. 5,000 bp) from the structural gene has been reported in sorghum through the cloning of a major QTL for aluminum tolerance [69]. Clearly, only a positional cloning approach is able to unequivocally highlight the role of noncoding regions in controlling the level of expression of a particular gene and the resulting phenotype. To what extent noncoding, long-distance enhancers might be involved in regulating the expression of quantitative traits is presently unknown. Notwithstanding the importance of this issue for a more complete understanding of the regulation of gene expression, this level of genetic dissection is certainly not required from a breeding standpoint, since both MAS and genetic engineering would still allow breeders to fully exploit the beneficial effects linked to either natural allelic variation or the ectopic expression of the structural locus encoding for the target trait.

Despite the clear advantages of association mapping on biparental linkage mapping (e.g., multiallelism, higher genetic variability and genetic resolution, no need to assemble a mapping population, shorter time required to identify relevant loci, etc.), a major limitation of the former is represented by the high rate of false positives (i.e., Type-I error rate), hence spurious association, due to the presence of hidden population structure among the accessions being evaluated [62]. An additional constraint to a more widespread utilization of association mapping for the dissection of physiologically complex traits may derive from factors other than statistical issues. For highly integrative and functionally complex traits such as yield, particularly under adverse conditions, association mapping may quickly lose its effectiveness as the level of functional complexity of the target trait increases. In this case, similar phenotypic values in different genotypes can result from the action of different gene networks and/or trait compensation (e.g., yield components), thus undermining the identification of significant marker-trait association across a broad range of genotypes such are those usually present in the panels used for association mapping. Although a similar limitation also pertains to a mapping population developed from the cross of two divergent lines, its relevance in the case of association mapping for complex traits is greatly increased by the much wider functional variability explored with association mapping. This is particularly the case whenever the investigated trait (e.g., yield under drought conditions) is strongly influenced by differences in phenology, mainly flowering time and/or plant height; in this case, the overwhelming effects on yield of phenological traits will inevitably overshadow the effects due to the action of loci controlling yield per se, i.e., irrespectively of flowering time and plant height.

Comparative QTL Mapping and Metanalysis

A major shortcoming in QTL mapping is the limited accuracy in identifying the most likely position of each single QTL on the chromosome. Unless highly isogenic materials are evaluated, the confidence interval in assigning a QTL is rarely shorter than 10 cM, an interval likely to contain several hundred genes. The availability of QTL data for two or more mapping populations of the same species allows for the comparison of the position of QTLs by means of a metanalysis carried out with dedicated software [70]. This, in turn, provides a better genetic resolution of the QTL interval and reduces the confidence interval around the peak of the LOD profile. This exercise is particularly useful when a reference map with hundreds of well-spaced markers is available and contains "anchor markers" (usually RFLPs, SSRs, and/or SNPs) also used to investigate other mapping populations of the same species. An additional advantage of a reference map is that it allows one to compare the map position of QTLs with that of mutants for the same trait, thus contributing relevant information for the identification of possible candidate genes causally affecting the investigated trait. Accordingly, Robertson [71] suggested that a mutant phenotype may be caused by an allele with a much more drastic effect in comparison to that of QTL alleles at the same locus, a hypothesis that has been validated in maize for a QTL for plant height colocalized with the mutant *dwarf3* [72]. These results indicate that no real boundary exists between Mendelian and quantitative genetics, while suggesting that loci can be classified in either category based upon the magnitude and heritability of the effect of the alleles being considered. It follows that the information provided by mutants is of great value for QTL studies and breeding applications.

Isogenic Materials for Mapping and Cloning QTLs

A valuable opportunity for investigating the effects of a particular QTL and eventually isolate the functionally polymorphic sequence responsible for its effects is offered by the analysis of pairs of isogenic materials (e.g., near isogenic lines: NILs) contrasted for the parental chromosome regions (usually ca. 10–30 cM long) present at the target QTL. NILs can be obtained through repeated selfings of F_3-F_5 individuals heterozygous at the QTL region prior to isolating the homozygotes for each one of the two parental segments carrying the functionally contrasting QTL alleles [73]. Alternatively, each parental line of the mapping population originally evaluated for discovering the QTL can be used as recurrent parent in a backcross scheme in which a single genotype heterozygous at the QTL in question is utilized as donor of the alternative QTL alleles; in this case, the congenic lines are identified as backcrossed-derived lines [74]. With NILs, it is thus possible to “mendelize” major QTLs characterized by a sizable additive effect. Unlike genome-wide QTL studies where more than 100–150 genotypes are usually screened, experiments conducted with NILs involve few genotypes (two as a minimum), thus allowing for a much more refined and detailed phenotypic evaluation of the effects of the QTL [74, 75]. However, it should always be appreciated that the results of NIL-based studies could to a certain extent be biased by the action of one or more closely linked genes affecting the investigated traits, a particularly likely event when the region flanking the QTL extends for several cM.

A more systematic search of QTLs is made possible with the use of a series of isogenic lines obtained through the introgression, via backcrossing, of a small portion (ca. 20–30 cM) of the genome of a donor line into a common recurrent line, usually an elite cultivar [76]. The final objective is to assemble a collection of so-called introgression library lines (ILLs; at least 70–80 or more lines for each cross), basically a collection of NILs, each one differing for the introgressed chromosome portion and collectively representing the entire donor genome [76]. A major advantage of ILLs is the rapid progress that they allow for the fine mapping and positional cloning of major QTLs [48, 77]. Besides the well-documented effectiveness of ILLs for the mapping and cloning of QTLs in tomato [77, 78], ILLs have been instrumental for mapping drought-adaptive QTLs in rice [79] and maize [163]. Once ILLs are made available and major loci for the target traits are identified, testing for epistasis becomes particularly feasible using a small number of genotypes, unlike with mapping populations, where an accurate testing for epistasis will require the evaluation of at least 200 families.

The availability of NILs for a major QTL is an important prerequisite for undertaking the cloning of the sequence underlying the trait being targeted. Besides contributing to a better understanding of the functional basis of quantitative traits [68, 80], QTL cloning provides an essential opportunity for more effectively mining and exploiting the allelic diversity present in germplasm collections [49, 82]. Recent advances in high-throughput profiling and sequencing of both the genome and transcriptome coupled with reverse-genetics approaches/platforms (e.g., collections of knockout mutants, TILLING, RNAi, etc.) have streamlined the procedures and markedly reduced the time required to identify the sequences governing variation in quantitative traits. Until now, the molecular dissection of a candidate locus has been prevailingly achieved through positional cloning and association mapping. Both approaches exploit LD to identify the most promising candidate gene(s) and benefit from the map information of candidate genes and mutants in the species under investigation and in closely related ones. As sequence information accumulates and our understanding of biochemical pathways improves, QTL cloning via the candidate-gene approach becomes an attractive alternative to positional cloning, particularly for traits underlined by a known metabolic pathway [83, 84].

Modeling QTL Effects

QTL-based modeling holds promise to allow for a more effective design of "molecular ideotypes" on the basis of estimated QTL effects for growth parameters of response curves to environmental factors revealed by exposing mapping populations to such environmental factors [85–87]. Additionally, crop modeling provides useful clues to unravel the genetic basis of G × E interactions and toward a better understanding of traits' plasticity [88], a feature of increasing importance in view of the effects on crop growth and yield due to the enhanced vagaries in weather conditions consequent to global warming. An accurate estimate of the consistency of QTL effects in a particular genetic background can be obtained through extensive testing of the genetic materials under different environmental conditions as to level of irrigation, nutrients, temperature, etc.

In maize, an ecophysiological model and QTL analysis have been integrated to investigate the genetic basis of leaf growth in response to drought and predict leaf elongation rate as a function of estimated QTL effects at varying air humidity, temperature, and soil water status (Tardieu 2003). QTLs with a limited QTL × E interaction and with a linear response to a particular environmental factor will provide more predictable opportunities to improve crops' performance through MAS. An important issue rarely addressed in view of the inherent difficulty in doing so from an experimental standpoint under field conditions is that crop performance is often constrained by more than one environmental factor (e.g., drought and heat) occurring simultaneously, a condition which greatly undermines the prediction of QTL effects, particularly when considering multiple QTLs.

Marker-Assisted Breeding to Improve Crop Performance

The improvement of crop performance through conventional breeding has for the most part been achieved with little or no knowledge of the genetic basis of the selected traits, particularly yield and its underlying morphophysiological determinants. The main obstacle to raising crop yield via conventional breeding by means of phenotypic selection is represented by the low heritability of yield, particularly under marginal conditions and low-input agriculture (e.g., low supply of nutrients and/or water). As an alternative to phenotypic selection, MAB can be applied to more effectively improve crop performance. The ultimate goal of MAB is to increase the cost-effectiveness of the selection gain per unit time. Although the costs entailed by MAB are still quite high when compared to conventional breeding practices, the sizable reduction in the time required to release an improved cultivar made possible through MAB can justify its application once agronomically valuable alleles at target loci (genes or QTLs) are identified. The convenience of adopting MAB to improve the efficiency of the selection process should be carefully evaluated on a case-by-case basis. The success of MAB will depend on the identification of the agronomically beneficial alleles at target loci, their effect in the different elite genetic backgrounds prevalently grown by farmers and their pyramiding in the correct combinations. MAB could thus be regarded as an extension and evolution of the so-called ideotype breeding, an approach based on phenotypic selection for an ideotype characterized by those morphophysiological features deemed necessary to maximize yield. As compared to ideotype breeding, MAB allows us to dissect the genetic basis of key traits and to piece back together the best alleles in a sort of molecular jigsaw puzzle, the main limitation being that only a very small number of the jigsaw tassels (i.e., genes and QTLs) have been identified. This approach, referred to as "breeding by design" [89], extends the concept of "graphical genotypes" first introduced by Young and Tanksley [90] to portray the parental origin and allelic contribution of each genotype on a genome-wide basis. Although a breeding-by-design approach is technically applicable to all major crops, its impact has been much more tangible for traits with a simple genetic control (e.g., quality, disease resistance; [91–95]) as compared to more complex quantitative traits, such as yield under adverse environmental conditions [60], a result mainly due to our rudimental understanding of the genetic basis of the latter category of traits, their interaction with environmental factors and, most importantly, the difficulty in predicting the phenotypic value of a new genotype tailored through MAB for several QTLs. Along this line, it should be underlined that the effects of QTL alleles for complex traits (e.g., yield) characterized by a large G × E interaction can drastically change according to the conditions (e.g., water availability along the crop life cycle) present in the environment being targeted.

The molecular profiles obtained with molecular markers provide the basic information required to identify the haplotype of each individual plant at a target locus. Haplotype profiling of collections of elite cultivars released during the past decades and derived from a limited number of founders (i.e., genotypes that in view of their positive features have been frequently used by breeders as parental lines) provides a means to identify the chromosome regions that have been preferentially retained throughout the breeding activities carried out during such time period. It is plausible to hypothesize that these chromosomal regions harbor loci (genes or QTLs) important for the selection of improved cultivars.

The strategies deployed to improve crop performance based on molecular information can be categorized according to the level of knowledge and understanding of the loci that underline the phenotypic traits under selection. While MAS and marker-assisted recurrent selection (MARS) during the past two decades have deployed allelic variation at mapped loci often characterized by a rather large effect on the phenotype, the new paradigm ushered in by genomic selection (GS) via high-throughput profiling has emphasized the selection of unmapped, uncharacterized loci with rather limited individual effects but with otherwise sizable effects when selected together. The next sections will critically analyze some of the main features of these rather different approaches that should not be regarded as antagonistic, but rather complementary.

Marker-Assisted Selection

Once loci are mapped and their effects characterized, the two most common applications of MAS in crop breeding are to (1) accelerate the backcross (BC) procedures required to transfer beneficial alleles at one or more loci into an elite cultivar and (2) facilitate the selection of one or more target traits within a segregating population. The former application is the one that so far has been most frequently adopted in breeding programs and is usually referred to as marker-assisted backcross (MABC). MAS has also been deployed frequently to create isogenic lines (e.g., NILs, introgression libraries, etc.). These materials are used to identify and map genes/QTLs and, as such, usually do not impact directly on the outcome of breeding practices and the release of improved cultivars.

As compared to the conventional BC procedure, MABC based on the use of markers uniformly spaced along the genome (ca. 20–25 cM apart) can save three to four BCs in recovering most of the genome of the recurrent parent, thus reducing the time required for the release via BC of the improved version of the recurrent parent [96]. The advantage is greater for the incorporation via BC of recessive resistance genes, the phenotypic detection of which is only possible for the homozygous individuals carrying recessive alleles at both loci. In this case, phenotypic selection takes twice longer as compared to dominant alleles, since a selfing generation is required after each BC for the phenotypic identification of the homozygous recessive resistant plants to be used for the next BC. The utilization of codominant markers (e.g., SSRs) allows for the identification of heterozygous plants carrying the resistance-encoding allele directly in F_1, thereby saving one generation for each BC cycle. During the past two decades, MABC has been routinely deployed by seed companies to introgress beneficial alleles from unadapted accessions (e.g., landraces or wild, sexually compatible relatives of crops) and particularly to introgress transgenes into elite materials [9, 97, 98]. At each generation, individuals heterozygous at the region flanking the target locus are identified based on the results of molecular profiling. In comparison to conventional BC, MABC provides additional, distinct advantages such as (1) avoiding the vagaries in phenotyping when the conditions do not allow an accurate classification of the progeny segregating for the target trait (e.g., absence of the pathogen when backcrossing an allele for resistance to the disease), (2) reducing the number of plants to be screened in each selection cycle, and (3) identifying plants with the shortest possible chromosome segment introgressed from the donor line. The latter factor is particularly important when the donor is a wild accession of the recurrent, elite line being backcrossed. In this case, the introgressed chromosome segment flanking the target locus is likely to contain many alleles with a detrimental effect on quality and yield. Therefore, it is necessary to select individuals with the shortest possible chromosomal fragment contributed by the donor parent. An additional benefit is when the phenotyping of the trait under transfer is expensive and/or cumbersome like in the case of genes affecting tolerance to diseases/pests that require artificial inoculation in order to correctly identify those plants carrying the tolerant alleles (e.g., resistance to nematodes; [99]). Other cases where MABC provides a distinct temporal advantage as compared to conventional procedures is when the phenotypic evaluation of the target trait is destructive or when the trait is expressed after flowering. Selection before flowering greatly reduces the number of plants to be selfed or crossed, thus reducing the operating costs, particularly with species with a long life cycle.

During backcrossing, different rates of recovery of the recipient genome are expected at the target region and the nontarget chromosomes. Because each BC reduces by half the percentage of the donor genome at nontarget regions, at least six or seven BCs are required for a satisfactory recovery (ca. 99%) of the recipient genome. However, the number of BCs is frequently higher due to residual linkage drag around the target locus and it is not uncommon that up to nine or ten BCs are implemented before the improved cultivar is finally released. Clearly, the longer the time required to complete the BC procedures, the lower the probability of success of the new cultivar, since other improved, competing cultivars will be released in the meantime. Simulation and practice have both shown that in a moderately sized population of a species with a relatively small genome (<500 million bp, such as rice) using more than two to three well-spaced markers per chromosome arm hardly brings any additional benefit. For a species with large chromosomes (e.g., wheat, ca. 16 billion bp), a larger number of markers in each chromosome are beneficial. With an increasing genome size, more independent recombination events are needed to reduce the contribution of the donor parent, which in turn requires a larger population size. To what extent the contribution of the donor parent should be reduced will depend on the type of alleles carried by such fragments and, most importantly, the genetic distance between the donor parent and the recurrent parent. Nowadays, the availability of large number of SNPs in most of the major crops facilitates the screening of the BC individuals to verify in great detail to what extent the genome of the donor parent has been retained.

Formulas are available to compute the level of concordance between the allelic state at the target locus and the flanking markers during the BC procedures [81]. These formulas values indicate that the level of control made possible with only one marker is insufficient to keep the risk of losing the target allele below 5% throughout five cycles of BC. Conversely, the level of control possible with two flanking markers is considerably higher even when the markers are not tightly linked to the target locus. If the BC procedure targets a QTL instead of a Mendelian locus, the uncertainty about the exact position of the sequence underlining the QTL introduces further complexity. Because the quantity of donor genes on the carrier chromosomes decreases much more slowly in comparison to the noncarrier chromosomes, after six BCs the majority of heterozygous loci with undesirable donor alleles will be on the carrier chromosome, with the vast majority included in the intact fragment flanking the target locus.

At the chromosomes not targeted by the BC procedure, it is expected that after “n” BCs, the probability that any locus remains heterozygous between the donor and the recipient is $(0.5)^n$, which means that each BC halves the residual level of heterozygosity. Consequently, six BCs ensure a level of similarity with the recurrent parent above 99%. Results in different species have shown that there may be a significant deviation from the 75% genomic portion of the recurrent parent expected in the BC_1 generation [100, 101], thus demonstrating the usefulness of genotype-based selection to identify plants with the highest possible portion of the genome from the recurrent parent.

Pyramiding Beneficial Alleles at Multiple Loci

The possibility to rapidly introgress and pyramid into existing cultivars a suite of beneficial alleles allows breeders to more quickly release improved cultivars to farmers. The best examples are in the area of disease resistance. Monogenic (Mendelian) resistance based on a single major gene is usually nondurable due to the high mutation rate in plant pathogens, which can lead to the selection of new virulent strains able to overcome the physiological barrier of an individual resistance gene. Consequently, the durability of disease resistance can be increased by screening for new sources of resistance followed by marker tagging of the relevant genes and their incorporation in elite cultivars. Pyramiding identifies the procedure for stacking the beneficial resistance alleles in a single line or cultivar, which provides a more durable resistance to pathogens as compared to monogenic resistance based on a single major gene. The advantage of pyramiding multiple alleles for resistance is particularly evident with diseases that require repeated inoculation and when phenotypic selection alone is too cumbersome and fails altogether to detect and combine multiple resistance genes in a single genotype.

Direct disease screening based on phenotypic observations is not always desirable due to a number of factors: quarantine restrictions, lack of routine screening methods and informative pathogen races for discriminating specific resistance genes, host escapes, and/or the inability to identify specific genes or gene combinations due to the occurrence of race or pathogen mixtures in the field. In these cases, MAS of race-specific genes offers a viable alternative for stacking beneficial alleles in improved genotypes which will eventually turn into novel cultivars characterized by more durable resistance to rapidly changing pathogen populations. Along this line, the constant changes in pathogen populations in different environments underline the potential value of previously defeated resistance genes. In this case, MAS offers the only practical solution to maintain such genes in current cultivars since they are masked by the epistatic effects of other resistance genes that are still effective.

In all major crops, the availability of markers tightly linked to resistance loci now allows breeders to tailor new cultivars with a suite of resistance genes able to enhance durable disease resistance to highly variable pathogens [102]. In broader terms, pyramiding is also implemented for combining beneficial alleles at loci (Mendelian or QTLs) that control traits other than disease resistance. In wheat, alleles at major loci that influence quality (e.g., semolina color, protein content, micronutrient concentration, etc.) and tolerance to abiotic stress (e.g., aluminum, boron, salinity, etc.) are routinely introgressed via MABC [94].

When multiple loci are targeted in a BC program, the minimum population size to be considered increases considerably and rapidly becomes a major limiting factor when more than three or four loci are targeted, a number that can be increased to five or six when Mendelian loci are considered. When the targeted loci are QTLs, the uncertainty of the exact location of each selected QTL adds further constraints and reduces the number of loci that can be selected with a population of manageable size. When different lines contribute the beneficial alleles, the easiest strategy is to cross them to produce recombinant progenies and select the desired individuals. Multiple crosses might be required to pyramid all the desired alleles in one single genotype. A more general framework and the underlying theory to optimize breeding schemes for gene pyramiding have been described [103].

Marker-Assisted Selection in a Segregating Population

MAS has been extensively used for the selection of single genes conferring tolerance to diseases/pests [91, 94, 102, 104–106]. Although early simulation studies suggested the effectiveness of MAS for the improvement of biparental populations segregating for moderately complex traits [107], the first applications of MAS in maize were disappointing [57, 108]. Sweet corn is the only exception, the main reason being its much narrower genetic basis as compared to maize used for feed production, a feature that increases the reliability of predicted gains from selection and extrapolation of the effects of different loci to different populations [109]. Another feature that makes the application of MAS particularly attractive in sweet corn is the high costs associated to conventional phenotyping, in view of the large amount of grain that needs to be processed in order to obtain an accurate estimate of the phenotypic values of the progeny to be selected. MAS applications have been more widespread in the private sector as compared to public institutions, most likely owing to a lack in the latter of the infrastructure required for an effective exploitation of MAS.

Notwithstanding the remarkable progress in identifying and in some cases cloning major loci regulating agronomically valuable traits [48, 49], more limited success has been reported for MAS of quantitative traits [110], mainly due to the difficulty in identifying major QTLs with a sufficiently large and stable effect for justifying their deployment via MAS. While true QTL × E interaction due to variable expression of a trait may cause lack of consistency in QTL detection particularly with traits characterized by low to moderate heritability, the interaction between a mapping population of small size – hence with limited power in QTL detection – with variable environments is probably an equally important factor causing inconsistency in QTL detection. This is particularly evident for the improvement of crop yield under drought conditions, one of the most difficult traits to improve not only via MAS [14, 60, 111–113] but also through conventional breeding.

Marker-Assisted Recurrent Selection

Although marker-assisted recurrent selection (MARS) was first proposed in the early 1990s [114], only recently its adoption has provided a tangible contribution to crop improvement, mostly due the difficulty in identifying multiple loci characterized by limited G × E interaction and reasonably consistent effects in different genetic backgrounds other than that in which they were originally identified. The goal of MARS is pretty much similar to that pursued in pyramiding alleles at multiple loci, i.e., accumulating the beneficial alleles at as many as possible, preferably all, loci being targeted. Pyramiding alleles at many loci (e.g., >10) is best achieved with a recurrent selection strategy [115]. In this case, simulation showed that with 50 QTLs and a population of 200 plants the frequency of favorable alleles reached 100% after ten cycles when markers cosegregated with the QTL (i.e., they coincided), but only 92% when the marker-QTL interval was equal to 5 cM, hence increasing the possibility of losing the desired QTL allele due to recombination. In practice, with a higher number of loci under selection the occurrence of plants carrying the desired ideal combination becomes increasingly unlikely and basically impossible when more than 20 loci are targeted simultaneously. This problem can be partially mitigated through successive cycles of crossing individuals carrying complementary combinations of the desired alleles [89]. This concept can be extrapolated to crosses with multiple parents.

MARS can start irrespectively of knowing the map position of the desired loci, which instead can be identified during the selection process. Simulation has clearly shown the superiority of MARS over phenotypic selection (from 5% to 20%), particularly when the selected population is highly heterozygous [116]. In maize, MARS has been applied rather extensively for improving relatively complex traits such as disease resistance, tolerance to abiotic stress, and also grain yield [111, 117–119].

The outcome of both MAS and MARS within a segregating population can be influenced by the genetic makeup of the targeted genetic background in terms of alleles present at other loci that interact epistatically with the target locus, an aspect which becomes particularly relevant for quantitative traits in view of the high number of loci involved in their control. Accordingly, since most evaluations of QTL effects and MAS strategies assume that QTLs act independently [55], it has been argued that MAS has little if any power over traditional phenotypic selection [46]. With maize as a model species, computer simulation showed that gene information is most useful in selection when few loci (<10) control the trait, while with many loci (>50) the least squares estimates of gene effects become imprecise. Based on these results, the typical reductionist approach pursued through QTL discovery strongly limits the outcome of MAS carried out for traits controlled by many QTLs [46].

Genomic Selection

In genomic selection (GS), genetic markers in number sufficient to cover the entire genome according to the level of LD are used so that most QTLs controlling the trait being selected are in LD with at least one neighboring marker. Unlike in MAS, in GS the individual plants are chosen without mapping the underlying QTLs that remain unknown along the entire process. Originally devised for animal breeding, only recently has GS been adopted for improving crop performance [120–122]. This was due to the fact that only in the past few years its application has become technically feasible in plants thanks to the introduction of SNP profiling with a level of genome saturation sufficient to detect the cumulative effects of the plethora of minor QTLs affecting quantitative traits which, on a single basis, will inevitably remain undetected in a biparental mapping population.

In GS, the breeding values of all the markers distributed across the genome are fitted as random effects in a linear model. The trait values are then predicted as the sum of the breeding values of each individual genotype across all the profiled markers and selection is based on these genome-wide predictions. A simulation study showed that across different numbers of QTLs (from 20 to 100) and levels of heritability, the response to GS was from 18% to 43% higher as compared to MARS. The number of markers that are used to predict the breeding values usually varies from a minimum of ca. 200 up to 500. A higher number of markers are required as the functional complexity of the targeted trait increases and LD decreases. Notably, GS is most effective for complex, low-heritable traits controlled by a large number of QTLs.

Implementation of GS is already having a major impact on the improvement of yield and other complex traits, mainly in the private sector where high-throughput infrastructures and robots allow for the routine creation and handling of millions of datapoints. Clearly, GS is not antagonistic to either MAS or MARS. Rather, they should be deployed in a complementary fashion on a case-by-case basis and according to the availability of mapped major QTLs, the accurate evaluation of their effect, and the frequency of the agronomically desirable alleles in the germplasm under selection.

Integrating Marker-Assisted Breeding in Conventional Breeding Projects

Among other factors, a broader application of MAB in conventional breeding projects will depend on the cost of molecular profiling [123, 124]. SNP markers are ideally suited for this role. In maize, the cost-effectiveness of MAS for the introgression of a single dominant allele into an elite line was compared with that of conventional breeding [125]. In this particular case, neither method showed clear superiority in terms of both cost and speed: Conventional breeding schemes were found to be less expensive while MAS-based breeding schemes were shown to be faster. High-throughput genotyping based on the scoring of markers that do not need the use of gels [126–128] coupled with quick DNA extraction protocols are needed to streamline MAS and lower its cost.

An important factor to be carefully considered prior to embarking in any MAS activity targeting specific loci is the robustness of the marker-locus association and their genetic distance. Clearly, the level of LD of the genetic materials used to investigate the genetic makeup of the target traits plays a pivotal role in determining the level of genetic resolution. Accordingly, biparental F_2 populations have the maximum amount of LD, hence the lowest level of genetic resolution. Although this feature is advantageous for the initial QTL mapping studies in view of the limited number of markers that are required, it clearly limits the accuracy of MAS and usually does not allow us to resolve tightly linked QTLs from pleiotropic ones [129]. This problem can be circumvented by deploying genetic materials that capture a higher recombinational level, either historically (e.g., panels of unrelated genotypes suitable for association mapping; [67, 130]) or through subsequent random matings of the individuals of the original mapping population [131]. Increasing the genetic resolution not only enhances the reliability of MAS but also reduces the list of the possible candidates, an important prerequisite in identifying the sequence responsible for the phenotype of interest. Therefore, prior to undertaking an association mapping study, it is important to acquire a good understanding of the LD patterns in the set of genetic materials to be evaluated. In fact, LD can be caused by factors other than linkage. Spurious associations in a collection of germplasm accessions can be due to LD between unlinked genomic regions (i.e., >50 cM apart) on the same chromosome and/or between genomic regions located on different chromosomes. Dedicated softwares are available to reduce the frequency of false-positive associations due to the bias introduced by preexisting population structure.

One of the most critical steps in any breeding program is the choice of suitable parental lines to create the new segregating populations that will undergo selection. Ideally, such parental lines will contribute beneficial alleles at the loci most critical for the target traits and, more in general, crop performance and its quality. Molecular profiling can contribute in two major ways to expedite the selection process and increase the response to selection. In autogamous crops, MAS is applied to choose the parental lines that are crossed to generate new mapping populations (mostly biparental) and then to select during the subsequent generations the recombinant progeny that carry the desired alleles at the targeted loci. In wheat, MAS is being deployed in a number of breeding programs both in the public and private sectors [94]. In particular, more than 30 traits have been targeted, mainly for disease resistance, quality, and abiotic stress tolerance. In allogamous crops (e.g., maize) where the populations used to extract new parental lines routinely undergo recurrent selection, MARS can be applied at each selection cycle to increase the frequency of the beneficial alleles within the population until the best performing alleles are fixed within the population and, as such, no longer require selection. By increasing the frequency of beneficial alleles in a breeding population, the probability of recovering a genotype with the combination of desired alleles is increased. As an example, increasing the favorable allele frequency from 0.50 to 0.96 will increase the probability of recovering the ideal genotype for 20 independent regions from one in a trillion to one in five [9]. This change in allele frequency will improve the mean performance for the selected trait of the population and any line derived from it. Breeders can deploy different MARS schemes depending on the selection model and the desired genetic structure (e.g., inbreeding level) of the population obtained after MARS. The MARS schemes require optimization for best managing field and laboratory resources, hence containing the costs, as well as for expediting the selection process, hence the accumulation of favorable allele frequency. When several traits and loci are targeted simultaneously, a multiple trait index is used to combine the values of each individual trait into a single index and different weights are assigned according to the perceived importance of each trait. The output of this process is an estimated genotypic value calculated for each progeny being considered for selection. MARS can also be applied to autogamous crops (e.g., soybean) in order to enhance the performance of the breeding populations used to select improved genotypes that will hopefully outperform the existing cultivars. As compared to conventional breeding practices, the outcome of MARS has clearly indicated its superiority for improving yield in maize, sunflower, and soybean [9]. Of utmost importance for the successful implementation of MARS is that breeders perform phenotypic selection on the lines per se that will be utilized for MARS. Additionally, phenotypic evaluation and selection among and within derived lines should continue after MARS.

Systematic profiling of parental lines is now routinely applied with a different level of genetic resolution, hence according to the level of LD of the target species. SSR profiling is rapidly being replaced by SNP profiling, much more effective than the former to define haplotype structure and much cheaper and amenable to high-throughput profiling. SNP platforms are particularly suited to the high-throughput profiling required by GS.

Once the template sequence of a crop becomes available, resequencing of lines can be used to obtain a far deeper understanding of their genomic architecture, allelic composition, and ultimate haplotype [132–134]. The spectacular reduction in cost that followed the introduction of second-generation sequencers makes resequencing of single genotypes a rather attractive and affordable option [135–137]. Additional progress in sequencing will further reduce the costs in as much direct resequencing of entire mapping populations may soon become more affordable than SNP profiling.

Mining Beneficial Alleles in Wild Relatives of Crops

As compared to their wild counterparts, the domestication bottleneck that all crops went through coupled with the strong selection first empirically practiced by farmers and then more systematically by modern breeders have markedly reduced the level of genetic variability within cultivated species, an aspect even more relevant for traits playing a substantial role in survival under natural conditions [82]. This limitation can be overcome through the implementation of advanced backcross QTL (AB-QTL) analysis [138], an approach that allows breeders to quickly discover and exploit beneficial QTL alleles present in wild germplasm but otherwise absent from elite germplasm. The AB-QTL approach relies on the evaluation of BC families between an elite cultivar utilized as recurrent parent and a donor accession, usually a wild species that is sexually compatible with the crop. Usually, QTL analysis is delayed until the BC_2 generation and after selection in BC_1 against features known to affect negatively yield (e.g., ear shattering in small-grain cereals). The effectiveness of the AB-QTL approach has been proven in tomato [138, 139], rice [140], and barley [141]. These results are encouraging for using AB-QTL as a germplasm enhancement strategy for identifying wild alleles capable of improving the yield of the related crop, particularly under low-input agriculture and marginal environments where wild alleles may prove more beneficial, particularly for yield per se and disease resistance. An essential prerequisite is that the introgression of such beneficial alleles should bear no negative consequences when crops are grown under more favorable and high-yielding conditions.

Wild relatives of crop species can contribute to the identification of novel alleles for agronomically relevant traits by focusing on those loci that molecular evidence indicates as having been targeted by selection during both domestication and modern breeding [142]. To this end, the comparison of the allelic diversity present in elite accessions, landraces, and the undomesticated wild relatives of each crop allows for the identification of loci devoid of genetic variation within the elite germplasm, most likely as a result of domestication and subsequent man-made selection. The underlying assumption is that the loss of genetic diversity observed from the wild parent to the cultivated crop highlights the strong man-made selection at loci that control the expression of agronomically important traits, particularly those relevant for adaptation to abiotic stress. Therefore, both this “diversity screen” approach and the AB-QTL approach allow for the identification of valuable loci which would otherwise go undetected due to a lack of allelic diversity in the cultivated gene pool. An additional advantage of the diversity screen approach is that it allows for the identification of candidate genes of potential agronomic importance even without prior knowledge of gene function.

Leveraging the “-Omics” Platforms

During the past decade, a number of technologically sophisticated platforms have become available to collect a large amount of data on the dynamics of the transcriptome, proteome, and metabolome. The availability of these “-omics” profiling data facilitates the identification of candidate genes and provides us with a more holistic picture of the molecular events characterizing functions at the cellular, organ, and plant levels and how these are influenced by environmental cues [84, 143–146].

Unlike from the classical QTL positional cloning approach in which an adequately large mapping population is basically “interrogated” in order to identify the genetic determinants of QTLs, the candidate-gene approach capitalizes on gathering experimental evidence to support and validate the causal role of a coding sequence (e.g., glutamine synthetase gene) in governing variation for the putative target trait (e.g., nitrogen-use efficiency). The major advantage of the candidate-gene approach is that it bypasses the tedious and expensive procedures required by positional cloning. Identifying suitable candidate genes and elucidating their function can be expedited by combining different approaches and high-throughput -omics platforms applied to target crops and/or to model species. From a technical standpoint, combining laser-capture microdissection with the -omics platforms offers an unprecedented level of functional resolution at the tissue level, down to a single-cell layer [145]. Among the different platforms available for the mass-scale profiling of the transcriptome, microarrays have been more frequently utilized to investigate the changes in gene expression, particularly in plants exposed to adverse conditions [147–150]. Nonetheless, microarray platforms are quickly being replaced by high-throughput transcriptome sequencing by means of second-generation sequencing platforms [151].

Additional information on the changes in cellular metabolism is provided by the profiling of the proteome [152] and metabolome [153, 154] that, as compared to the transcriptome, are functionally closer to the phenotype, thus reporting also on variability due to posttranscriptional and posttranslational regulation. However, it should be appreciated that both proteomics and metabolomics report changes for a rather limited portion (ca. 5%) of the expressed genes; additionally, proteomics is often unable to detect the changes in gene products (e.g., transcription factors) that despite their low level are more likely to play an important role in pivotal functions (e.g., signal transduction in response to biotic and abiotic stress) and consequently, to underline QTLs.

Metabolome profiling can also be used to identify loci regulating the level of a particular metabolite and verify its coincidence with QTLs for yield and/or genes involved in metabolic pathways. With the present technology, up to ca. 2,000 different metabolites can be profiled in a single sample [155]. In maize, QTLs for invertase activity have been identified in a population subjected to drought stress [156]. The number of QTLs for invertase activity detected under drought was more than twice the number detected under well-watered conditions, an indirect indication of the important role of this enzyme under drought conditions. One QTL common to both treatments was located near *Ivr2*, an invertase-encoding gene. The colocation reported between the activities of three enzymes (invertase, sucrose-P synthase, and ADP-glucose pyrophosphorylase) involved in sucrose and starch metabolism and a corresponding structural gene suggests its role as a candidate gene for explaining part of the variability in enzyme activity [157]. These studies indicate that invertase activity is an important limiting factor for grain yield in maize exposed to drought during the reproductive phase [158].

The candidate-gene approach is particularly effective when a clear cause-effect relationship can be unequivocally established between the gene product and the target trait. An example of this approach is the cloning of a QTL for cell-wall beta-glucans in barley grains based on a synteny analysis between barley and rice that revealed the presence in the syntenic portion of the rice genome of a cellulose synthase-like *CslF* gene that genetic engineering unequivocally showed to influence beta-glucans content in barley grains as well as in other species, including also *Arabidopsis* [83]. This notwithstanding, identifying suitable candidates for functionally complex traits such as yield and yield components is a much more daunting undertaking given the large number of genes that influence these traits.

Future Directions

The first comprehensive report of DNA-based markers (RFLPs; [20]) in a crop species was published in 1986. Since then, an almost countless number of studies have shed light on the genetic control of plant growth and functions, and, most importantly crop yield. One clear take-home message that has emerged from these studies is the existence of a continuum between Mendelian and quantitative traits that will eventually help in identifying the functional polymorphisms, either of genetic or epigenetic origin that underlie quantitative trait variation. In this respect, QTL cloning will become a more routine and easier practice thanks also to the massive resequencing of mutant collections. This, in turn, will facilitate the identification of the best performing QTL alleles, their pyramiding through MAS, and the identification of novel alleles via TILLING [159] or by means of site-directed mutagenesis at the key functional domains of the encoded proteins. It is under this QTL cloning paradigm that the molecular basis of quantitative traits will be dissected in order to advance our understanding of the genetic makeup of this category of traits and to more accurately tailor crop morphology and productivity with beneficial alleles.

From an applicative standpoint, although conventional selection based on phenotypic evaluation will likely remain the mainstay for most breeding programs, particularly in the public domain, MAB and its applications will increasingly be adopted and will in some cases become prevalent as compared to conventional practices. As the twenty-first century unfolds, a multitude of genomics and postgenomics platforms are at hand to expand our understanding of the genetic basis of crop performance and to improve the efficiency of selection procedures for the release of new, improved cultivars. Resequencing will revolutionize the way breeders deal with their germplasm and will provide unsurpassed opportunities for a deeper mining of allelic diversity and harnessing its full potential. Nonetheless, our understanding of the functional basis of yield and other quantitative traits is likely to remain rudimental. The elusive nature of the QTLs that govern yield and yield stability is a formidable hurdle toward a more effective selection targeting specific loci and a better understanding of quantitative traits. Notably, GS can and will be applied irrespective of our degree of understanding of the genetic architecture of quantitative traits. Importantly, MAS and GS should be considered as complementary rather than alternative approaches, the utilization of which should be determined on a case-by-case basis. Bioinformatics and user-friendly databases will play a pivotal role for handling and managing the deluge of data produced by the molecular and phenotypic platforms.

In terms of experimental materials utilized for QTL studies, a growing attention will be devoted to the exploitation of multiparental crosses and mini-core collections of germplasm accessions with varying LD levels. In the mapping populations so far utilized for QTL discovery, most QTLs go undetected owing to the small size of the population, the presence of functionally monomorphic alleles and the small effects of many of such QTLs. Along this line, nested-association mapping (NAM) populations provide an interesting option to take advantage of both biparental (linkage) mapping and association mapping [160]. On a finer scale, high-throughput proteome and metabolome profiling will accelerate the identification of the causative mechanisms contributing to adaptive responses to adverse environmental conditions (e.g., drought, flooding, heat, etc.) whose frequency and intensity are expected to increase due to global warming. Nonetheless, the deluge of information originated through the molecular approaches and the -omics platforms will not automatically translate into novel cultivars. A "systems biology"-like approach will be instrumental for optimizing the accurate integration and exploitation in breeding terms of all the -omics information.

From an applicative standpoint, accurate phenotyping often remains the main limiting factor for identifying novel loci [161]. Semiautomated, high-throughput phenotyping under both controlled conditions and in the field promises to streamline gene discovery and narrowing the genotype-phenotype gap that hampers a more widespread deployment of MAB in crop improvement [87]. Along this line, it is important to emphasize that any molecular approach aiming to discover genes/QTLs and test their effects should preferably be carried out in an experimental context whose results are as relevant as possible and readily applicable to the conditions prevailing in farmers' fields [150]. An effective exploitation of genomics approaches to enhance crop performance will depend on their integration with conventional breeding. Although it is not possible to predict to what extent and how quickly the latter will be replaced by MAB, the future release of improved cultivars will be expedited and made more cost effective through a systematic marker-based manipulation of the loci that govern crop performance and the desired features targeted by breeders.

Bibliography

Primary Literature

1. Borlaug NE, Dowswell CR (2005) Feeding a world of ten billion people: a 21st century challenge. In: Tuberosa R, Phillips RL, Gale M (eds) In the wake of the double helix: from the green revolution to the gene revolution, proceedings of the international congress, 27–31 May 2003, Bologna, pp 3–23
2. Borlaug NE (2007) Sixty-two years of fighting hunger: personal recollections. Euphytica 157:287–297
3. Duvick DN (2005) The contribution of breeding to yield advances in maize (*Zea mays* L.). Adv Agron 86:83–145
4. Tester M, Langridge P (2010) Breeding technologies to increase crop production in a changing world. Science 327:818–822
5. Lande R, Thompson R (1990) Efficiency of marker-assisted selection in the improvement of quantitative traits. Genetics 124:743–756
6. Lee M (1995) DNA markers and plant breeding programs. Adv Agron 55:265–344
7. Morgante M, Salamini F (2003) From plant genomics to breeding practice. Curr Opin Biotechnol 14:214–219
8. Varshney RK, Graner A, Sorrells ME (2005) Genomics-assisted breeding for crop improvement. Trends Plant Sci 10:621–630
9. Eathington SR, Crosbie TM, Edwards MD, Reiter R, Bull JK (2007) Molecular markers in a commercial breeding program. Crop Sci 47:S154–S163
10. Yano M, Tuberosa R (2009) Genome studies and molecular genetics-from sequence to crops: genomics comes of age. Curr Opin Plant Biol 12:103–106
11. Flavell R (2010) Knowledge and technologies for sustainable intensification of food production. New Biotechnol 27: 505–516
12. Wei XJ, Liu LL, Xu JF, Jiang L, Zhang WW, Wang JK et al (2010) Breeding strategies for optimum heading date using genotypic information in rice. Mol Breeding 25:287–298
13. Schneeberger K, Weigel D (2011) Fast-forward genetics enabled by new sequencing technologies. Trends Plant Sci 16:282–288
14. Xu YB, Crouch JH (2008) Marker-assisted selection in plant breeding: from publications to practice. Crop Sci 48:391–407
15. Raju NL, Gnanesh BN, Lekha P, Jayashree B, Pande S, Hiremath PJ et al (2010) The first set of EST resource for gene discovery

and marker development in pigeonpea (*Cajanus cajan* L.). BMC Plant Biol 10:45

16. Shinozaki K, Yamaguchi-Shinozaki K (2007) Gene networks involved in drought stress response and tolerance. J Exp Bot 58:221–227
17. Fujii H, Zhu JK (2009) Arabidopsis mutant deficient in 3 abscisic acid-activated protein kinases reveals critical roles in growth, reproduction, and stress. Proc Natl Acad Sci USA 106:8380–8385
18. Yoshida T, Fujita Y, Sayama H, Kidokoro S, Maruyama K, Mizoi J et al (2010) AREB1, AREB2, and ABF3 are master transcription factors that cooperatively regulate ABRE-dependent ABA signaling involved in drought stress tolerance and require ABA for full activation. Plant J 61:672–685
19. Service RF (2006) Gene sequencing – the race for the $1000 genome. Science 311:1544–1546
20. Helentjaris T, Slocum M, Wright S, Schaefer A, Nienhuis J (1986) Construction of genetic linkage maps in maize and tomato using restriction fragment length polymorphisms. Theor Appl Genet 72:761–769
21. Tanksley SD, Young ND, Paterson AH, Bonierbale MW (1989) RFLP mapping in plant breeding: new tools for an old science. BioTechnology 7:257–264
22. Vos P, Hogers R, Bleeker M, Reijans M, Vandelee T, Hornes M et al (1995) AFLP: a new technique for DNA fingerprinting. Nucleic Acids Res 23:4407–4414
23. Senior ML, Chin ECL, Lee M, Smith JSC, Stuber CW (1996) Simple sequence repeat markers developed from maize sequences found in the GENBANK database: map construction. Crop Sci 36:1676–1683
24. Kilian A (2005) The fast and the cheap: SNP and DArT-based whole genome profiling for crop improvement. In: Tuberosa R, Gale M, Phillips RL (eds) In the wake of the double helix: from the green revolution to the gene revolution, proceedings of the international congress, 27–31 May 2003, Bologna, pp 443–461
25. Rafalski A (2002) Applications of single nucleotide polymorphisms in crop genetics. Curr Opin Plant Biol 5:94–100
26. Barbazuk WB, Emrich SJ, Chen HD, Li L, Schnable PS (2007) SNP discovery via 454 transcriptome sequencing. Plant J 51:910–918
27. Akhunov E, Nicolet C, Dvorak J (2009) Single nucleotide polymorphism genotyping in polyploid wheat with the Illumina GoldenGate assay. Theor Appl Genet 119:507–517
28. Close TJ, Bhat PR, Lonardi S, Wu YH, Rostoks N, Ramsay L et al (2009) Development and implementation of high-throughput SNP genotyping in barley. BMC Genomics 10:582
29. McCouch SR, Zhao KY, Wright M, Tung CW, Ebana K, Thomson M et al (2010) Development of genome-wide SNP assays for rice. Breeding Sci 60:524–535
30. Bentley DR (2006) Whole-genome re-sequencing. Curr Opin Genet Dev 16:545–552
31. Nordborg M, Weigel D (2008) Next-generation genetics in plants. Nature 456:720–723
32. Schnable PS, Ware D, Fulton RS, Stein JC, Wei FS, Pasternak S et al (2009) The B73 maize genome: complexity, diversity, and dynamics. Science 326:1112–1115
33. Varshney RK, Nayak SN, May GD, Jackson SA (2009) Next-generation sequencing technologies and their implications for crop genetics and breeding. Trends Biotechnol 27:522–530
34. Deschamps S, Campbell MA (2010) Utilization of next-generation sequencing platforms in plant genomics and genetic variant discovery. Mol Breeding 25:553–570
35. Moore G, Foote T, Helentjaris T, Devos K, Kurata N, Gale M (1995) Was there a single ancestral cereal chromosome? Trends Genet 11:81–82
36. Gale MD, Devos KM (1998) Plant comparative genetics after 10 years. Science 282:656–659
37. Bolot S, Abrouk M, Masood-Quraishi U, Stein N, Messing J, Feuillet C et al (2009) The 'inner circle' of the cereal genomes. Curr Opin Plant Biol 12:119–125
38. Falconer DS (1981) Introduction to quantitative genetics. Longman, London
39. Geldermann H (1975) Investigation on inheritance of quantitative characters in animals by gene markers. I. Methods. Theor Appl Genet 46:319–330
40. Sax K (1923) The association of size differences with seed-coat patterns and pigmentation in *Phaseolus vulgaris*. Genetics 8:552–560
41. Edwards MD, Stuber CW, Wendel JF (1987) Molecular-marker-facilitated investigations of quantitative-trait loci in maize. I. Numbers, genomic distribution and types of gene-action. Genetics 116:113–125
42. Stuber CW, Edwards MD, Wendel JF (1987) Molecular marker facilitated investigations of quantitative trait loci in maize. II. Factors influencing yield and its component traits. Crop Sci 27:639–648
43. Tanksley SD (1993) Mapping polygenes. Annu Rev Genet 27:205–233
44. Beavis WD (1998) QTL analysis: power, precision, and accuracy. In: Molecular dissection of complex traits. CRC Press, Boca Raton, pp 145–162
45. Ribaut JM, Hoisington D (1998) Marker-assisted selection: new tools and strategies. Trends Plant Sci 3:236–239
46. Bernardo R (2001) What if we knew all the genes for a quantitative trait in hybrid crops? Crop Sci 41:1–4
47. Bernardo R (2008) Molecular markers and selection for complex traits in plants: learning from the last 20 years. Crop Sci 48:1649–1664
48. Frary A, Nesbitt TC, Grandillo S, Knaap E, Cong B, Liu J et al (2000) fw2.2: a quantitative trait locus key to the evolution of tomato fruit size. Science 289:85–88
49. Salvi S, Tuberosa R (2007) Cloning QTLs in Plants. In: Varshney RK, Tuberosa R (eds) Genomics-assisted crop improvement-volume 1: genomics approaches and platforms. Springer, Dordrecht, pp 207–226
50. Churchill GA, Doerge RW (1994) Empirical threshold values for quantitative trait mapping. Genetics 138:963–971

51. Lander ES, Botstein D (1989) Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics 121:185–199
52. Darvasi A, Weinreb A, Minke V, Weller JI, Soller M (1993) Detecting marker-QTL linkage and estimating QTL gene effect and map location using a saturated genetic map. Genetics 134:943–951
53. Kao CH, Zeng ZB, Teasdale RD (1999) Multiple interval mapping for quantitative trait loci. Genetics 152:1203–1216
54. Doerge RW (2002) Mapping and analysis of quantitative trait loci in experimental populations. Nat Rev Genet 3:43–52
55. Podlich DW, Winkler CR, Cooper M (2004) Mapping as you go: an effective approach for marker-assisted selection of complex traits. Crop Sci 44:1560–1571
56. Beavis WD (1994) The power and deceit of QTL experiments: lessons from comparative QTL studies. In: Proceedings of the forty-ninth annual corn and sorghum research conference, Washington, DC, pp 250–266
57. Openshaw S, Frascaroli E (1997) QTL detection and marker-assisted selection for complex traits in maize. In: Annual corn and sorghum research conference American seed trade association. Washington, DC, pp 44–53
58. Melchinger AE, Utz HF, Schon CC (1998) Quantitative trait locus (QTL) mapping using different testers and independent population samples in maize reveals low power of QTL detection and large bias in estimates of QTL effects. Genetics 149:383–403
59. Abiola O, Angel JM, Avner P, Bachmanov AA, Belknap JK, Bennett B et al (2003) The nature and identification of quantitative trait loci: a community's view. Nat Rev Genet 4:911–916
60. Collins NC, Tardieu F, Tuberosa R (2008) Quantitative trait loci and crop performance under abiotic stress: where do we stand? Plant Physiol 147:469–486
61. Buckler ESI, Thornsberry JM (2002) Plant molecular diversity and applications to genomics. Curr Opin Plant Biol 5:107–111
62. Ersoz ES, Yu J, Buckler ES (2007) Applications of linkage disequilibrium and association mapping. In: Varshney RK, Tuberosa R (eds) Genomics-assisted crop improvement-volume 1: genomics approaches and platforms. Springer, Dordrecht, pp 97–120
63. Clark RM (2010) Genome-wide association studies coming of age in rice. Nat Genet 42:926–927
64. Rafalski A (2010) Association genetics in crop improvement. Curr Opin Plant Biol 13:174–180
65. Maccaferri M, Sanguineti MC, Demontis A, El-Ahmed A, del Moral LG, Maalouf F et al (2011) Association mapping in durum wheat grown across a broad range of water regimes. J Exp Bot 62:409–438
66. Buntjer JB, Sorensen AP, Peleman JD (2005) Haplotype diversity: the link between statistical and biological association. Trends Plant Sci 10:466–471
67. Buckler ES, Holland JB, Bradbury PJ, Acharya CB, Brown PJ, Browne C et al (2009) The genetic architecture of maize flowering time. Science 325:714–718
68. Salvi S, Sponza G, Morgante M, Tomes D, Niu X, Fengler KA et al (2007) Conserved noncoding genomic sequences associated with a flowering-time quantitative trait locus in maize. Proc Natl Acad Sci USA 104:11376–11381
69. Magalhaes JV, Liu J, Guimaraes CT, Lana UGP, Alves VMC, Wang YH et al (2007) A gene in the multidrug and toxic compound extrusion (MATE) family confers aluminum tolerance in sorghum. Nat Genet 39:1156–1161
70. Salvi S, Castelletti S, Tuberosa R (2009) An updated consensus map for flowering time QTLs in maize. Maydica 54:501–512
71. Robertson DS (1985) A possible technique for isolating genic DNA for quantitative traits in plants. J Theor Biol 117:1–10
72. Touzet P, Winkler RG, Helentjaris T (1995) Combined genetic and physiological analysis of a locus contributing to quantitative variation. Theor Appl Genet 91:200–205
73. Tuinstra MR, Ejeta G, Goldsbrough PB (1997) Heterogeneous inbred family (HIF) analysis: a method for developing near-isogenic lines that differ at quantitative trait loci. Theor Appl Genet 95:1005–1011
74. Landi P, Sanguineti MC, Salvi S, Giuliani S, Bellotti M, Maccaferri M et al (2005) Validation and characterization of a major QTL affecting leaf ABA concentration in maize. Mol Breeding 15:291–303
75. Landi P, Giuliani S, Salvi S, Ferri M, Tuberosa R, Sanguineti MC (2010) Characterization of root-yield-1.06, a major constitutive QTL for root and agronomic traits in maize across water regimes. J Exp Bot 61:3553–3562
76. Zamir D (2001) Improving plant breeding with exotic genetic libraries. Nat Rev Genet 2:983–989
77. Paran I, Zamir D (2003) Quantitative traits in plants: beyond the QTL. Trends Genet 19:303–306
78. Eshed Y, Zamir D (1995) An introgression line population of *Lycopersicon pennellii* in the cultivated tomato enables the identification and fine mapping of yield-associated QTL. Genetics 141:1147–1162
79. Li ZK, Fu BY, Gao YM, Xu JL, Ali J, Lafitte HR et al (2005) Genome-wide introgression lines and their use in genetic and molecular dissection of complex phenotypes in rice (*Oryza sativa* L.). Plant Mol Biol 59:33–52
80. Takahashi Y, Shomura A, Sasaki T, Yano M (2001) *Hd6*, a rice quantitative trait locus involved in photoperiod sensitivity, encodes the a subunit of protein kinase CK2. Proc Natl Acad Sci USA 98:7922–7927
81. Hospital F (2003) Marker-assisted breeding. In: Newbury HI (ed.) Plant molecular breeding. Blackwell Publishing and CRC Press, Oxford and Boca Raton, pp 30–59
82. Feuillet C, Langridge P, Waugh R (2008) Cereal breeding takes a walk on the wild side. Trends Genet 24:24–32
83. Burton RA, Wilson SM, Hrmova M, Harvey AJ, Shirley NJ, Stone BA et al (2006) Cellulose synthase-like *CslF* genes mediate the synthesis of cell wall (1,3;1,4)-β-D-glucans. Science 311:1940–1942

84. Belo A, Zheng PZ, Luck S, Shen B, Meyer DJ, Li BL et al (2008) Whole genome scan detects an allelic variant of *fad2* associated with increased oleic acid levels in maize. Mol Genet Genomics 279:1–10
85. Yin XY, Stam P, Kropff MJ, Schapendonk AH (2003) Crop modeling, QTL mapping, and their complementary role in plant breeding. Agron J 95:90–98
86. Hammer G, Cooper M, Tardieu F, Welch S, Walsh B, van Eeuwijk F et al (2006) Models for navigating biological complexity in breeding improved crop plants. Trends Plant Sci 11:587–593
87. Tardieu F, Tuberosa R (2010) Dissection and modeling of abiotic stress tolerance in plants. Curr Opin Plant Biol 13:206–212
88. Chapman SC (2008) Use of crop models to understand genotype by environment interactions for drought in real-world and simulated plant breeding trials. Euphytica 161:195–208
89. Peleman JD, van der Voort JR (2003) Breeding by design. Trends Plant Sci 8:330–334
90. Young ND, Tanksley SD (1989) Restriction fragment length polymorphism maps and the concept of graphical genotypes. Theor Appl Genet 77:95–101
91. Young ND (1999) A cautiously optimistic vision for marker-assisted breeding. Mol Breeding 5:505–510
92. Willcox MC, Khairallah MM, Bergvinson D, Crossa J, Deutsch JA, Edmeades GO et al (2002) Selection for resistance to southwestern corn borer using marker-assisted and conventional backcrossing. Crop Sci 42:1516–1528
93. Sørensen AP, Stuurman J, van der Voort JR, Peleman J (2007) Molecular breeding: maximizing the exploitation of genetic. In: Varshney RK, Tuberosa R (eds) Genomics-assisted crop improvement-volume 1: genomics approaches and platforms. Springer, Dordrecht, pp 31–56
94. Gupta PK, Langridge P, Mir RR (2010) Marker-assisted wheat breeding: present status and future possibilities. Mol Breeding 26:145–161
95. Xue SL, Li GQ, Jia HY, Lin F, Cao Y, Xu F et al (2010) Marker-assisted development and evaluation of near-isogenic lines for scab resistance QTLs of wheat. Mol Breeding 25:397–405
96. Randhawa HS, Mutti JS, Kidwell K, Morris CF, Chen XM, Gill KS (2009) Rapid and targeted introgression of genes into popular wheat cultivars using marker-assisted background selection. PLoS One 4:e5752
97. Hospital F, Charcosset A (1997) Marker-assisted introgression of quantitative trait loci. Genetics 147:1469–1485
98. Concibido VC, La Vallee B, McLaird P, Pineda N, Meyer J, Hummel L et al (2003) Introgression of a quantitative trait locus for yield from *Glycine soja* into commercial soybean cultivars. Theor Appl Genet 106:575–582
99. Zwart RS, Thompson JP, Milgate AW, Bansal UK, Williamson PM, Raman H et al (2010) QTL mapping of multiple foliar disease and root-lesion nematode resistances in wheat. Mol Breeding 26:107–124
100. Frisch M, Bohn M, Melchinger AE (1999) Comparison of selection strategies for marker-assisted backcrossing of a gene. Crop Sci 39:1295–1301
101. Frisch M, Bohn M, Melchinger AE (1999) Minimum sample size and optimal positioning of flanking markers in marker-assisted backcrossing for transfer of a target gene. Crop Sci 39:967–975
102. Maccaferri M, Mantovani P, Tuberosa R, DeAmbrogio E, Giuliani S, Demontis A et al (2008) A major QTL for durable leaf rust resistance widely exploited in durum wheat breeding programs maps on the distal region of chromosome arm 7BL. Theor Appl Genet 117:1225–1240
103. Servin B, Martin OC, Mezard M, Hospital F (2004) Toward a theory of marker-assisted gene pyramiding. Genetics 168: 513–523
104. Knapp SJ (1998) Marker-assisted selection as a strategy for increasing the probability of selecting superior genotypes. Crop Sci 38:1164–1174
105. Gebhardt C, Bellin D, Henselewski H, Lehmann W, Schwarzfischer J, Valkonen JPT (2006) Marker-assisted combination of major genes for pathogen resistance in potato. Theor Appl Genet 112:1458–1464
106. Ejeta G, Knoll JE (2007) Marker-assisted selection in sorghum. In: Varshney RK, Tuberosa R (eds) Genomics-assisted crop improvement-volume 2: genomics applications in crops. Springer, Dordrecht, pp 187–206
107. Edwards MD, Johnson L (1994a) RFLPs for rapid recurrent selection. In: Proceedings joint plant breeding symposium series. American Society Horticulture and Crop Science Society America, Corvallis
108. Moreau L, Charcosset A, Gallais A (2004) Experimental evaluation of several cycles of marker-assisted selection in maize. Euphytica 137:111–118
109. Yousef GG, Juvik JA (2001) Comparison of phenotypic and marker-assisted selection for quantitative traits in sweet corn. Crop Sci 41:645–655
110. Chen LM, Zhao ZG, Liu X, Liu LL, Jiang L, Liu SJ et al (2011) Marker-assisted breeding of a photoperiod-sensitive male sterile *japonica* rice with high cross-compatibility with *indica* rice. Mol Breeding 27:247–258
111. Ribaut JM, Ragot M (2007) Marker-assisted selection to improve drought adaptation in maize: the backcross approach, perspectives, limitations, and alternatives. J Exp Bot 58:351–360
112. Reynolds M, Tuberosa R (2008) Translational research impacting on crop productivity in drought-prone environments. Curr Opin Plant Biol 11:171–179
113. Herve P, Serraj R (2009) Gene technology and drought: a simple solution for a complex trait? Afr J Biotechnol 8:1740–1749
114. Edwards MD, Page NJ (1994) Evaluation of marker-assisted selection through computer simulation. Theor Appl Genet 88:376–382
115. Charmet G, Robert N, Perretant MR, Gay G, Sourdille P, Groos C et al (1999) Marker-assisted recurrent selection for cumulating additive and interactive QTLs in recombinant inbred lines. Theor Appl Genet 99:1143–1148
116. van Berloo R, Stam P (2001) Simultaneous marker-assisted selection for multiple traits in autogamous crops. Theor Appl Genet 102:1107–1112

117. Eathington SR, Dudley JW, Rufener GK (1997) Usefulness of marker-QTL associations in early generation selection. Crop Sci 37:1686–1693
118. Eathington SR (2005) Practical applications of molecular technology in the development of commercial maize hybrids. In: Proceedings of the 60th annual corn and sorghum seed research conferences. American Seed Trade Association, Washington, DC
119. Crosbie T, Eathington S, Johnson G, Edwards M, Reiter R, Stark S, et al (2006) Plant breeding: past, present, and future. In: Lamkey KR, Lee M (eds) Plant breeding: the Arnel R. Hallauer international symposium. Blackwell, Ames, pp 3–50
120. Heffner EL, Sorrells ME, Jannink JL (2009) Genomic selection for crop improvement. Crop Sci 49:1–12
121. Bernardo R (2010) Genomewide selection with minimal crossing in self-pollinated crops. Crop Sci 50:624–627
122. Lorenz AJ, Chao SM, Asoro FG, Heffner EL, Hayashi T, Iwata H et al (2011) Genomic selection in plant breeding: knowledge and prospects. Adv Agron 110:77–123
123. Moreau L, Lemarie S, Charcosset A, Gallais A (2000) Economic efficiency of one cycle of marker-assisted selection. Crop Sci 40:329–337
124. Kuchel H, Fox R, Reinheimer J, Mosionek L, Willey N, Bariana H et al (2007) The successful application of a marker-assisted wheat breeding strategy. Mol Breeding 20:295–308
125. Morris M, Dreher K, Ribaut JM, Khairallah M (2003) Money matters (II): costs of maize inbred line conversion schemes at CIMMYT using conventional and marker-assisted selection. Mol Breeding 11:235–247
126. Salvi S, Tuberosa R, Phillips RL (2001) Development of PCR-based assays for allelic discrimination in maize by using the 5′-nuclease procedure. Mol Breeding 8:169–176
127. Canovas A, Rincon G, Islas-Trejo A, Wickramasinghe S, Medrano JF (2010) SNP discovery in the bovine milk transcriptome using RNA-Seq technology. Mamm Genome 21:592–598
128. Mondini L, Nachit MM, Porceddu E, Pagnotta MA (2011) HRM technology for the identification and characterization of INDEL and SNP mutations in genes involved in drought and salt tolerance of durum wheat. Plant Genetic Resources-Characterization and Utilization 9:166–169
129. Graham GI, Wolff DW, Stuber CW (1997) Characterization of a yield quantitative trait locus on chromosome five of maize by fine mapping. Crop Sci 37:1601–1610
130. Darvasi A, Soller M (1995) Advanced intercross lines, an experimental population for fine genetic-mapping. Genetics 141:1199–1207
131. Lee M, Sharopova N, Beavis WD, Grant D, Katt M, Blair D et al (2002) Expanding the genetic map of maize with the intermated B73 × Mo17 (IBM) population. Plant Mol Biol 48:453–461
132. Atwell S, Huang YS, Vilhjalmsson BJ, Willems G, Horton M, Li Y et al (2010) Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. Nature 465:627–631
133. Bergelson J, Roux F (2010) Towards identifying genes underlying ecologically relevant traits in *Arabidopsis thaliana*. Nat Rev Genet 11:867–879
134. Huang XH, Wei XH, Sang T, Zhao QA, Feng Q, Zhao Y et al (2010) Genome-wide association studies of 14 agronomic traits in rice landraces. Nat Genet 42:961–976
135. Gupta PK (2008) Ultrafast and low-cost DNA sequencing methods for applied genomics research. Proc Natl Acad Sci India 78:91–102
136. Huang XH, Feng Q, Qian Q, Zhao Q, Wang L, Wang AH et al (2009) High-throughput genotyping by whole-genome resequencing. Genome Res 19:1068–1076
137. Lam HM, Xu X, Liu X, Chen WB, Yang GH, Wong FL et al (2010) Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. Nat Genet 42:1053–1059
138. Tanksley SD, Grandillo S, Fulton TM, Zamir D, Eshed Y, Petiard V et al (1996) Advanced backcross QTL analysis in a cross between an elite processing line of tomato and its wild relative *L. pimpinellifolium*. Theor Appl Genet 92: 213–224
139. Bernacchi D, Beck-Bunn T, Eshed Y, Inai S, Lopez J, Petiard V et al (1998) Advanced backcross QTL analysis of tomato. II. Evaluation of near-isogenic lines carrying single-donor introgressions for desirable wild QTL-alleles derived from *Lycopersicon hirsutum* and *L. pimpinellifolium*. Theor Appl Genet 97:170–180
140. Xiao J, Li J, Grandillo S, Ahn SN, Yuan L, Tanksley SD et al (1998) Identification of trait-improving quantitative trait loci alleles from a wild rice relative, *Oryza rufipogon*. Genetics 150:899–909
141. Talame V, Sanguineti MC, Chiapparino E, Bahri H, Ben Salem M, Forster BP et al (2004) Identification of *Hordeum spontaneum* QTL alleles improving field performance of barley grown under rainfed conditions. Ann Appl Biol 144:309–319
142. Yamasaki M, Tenaillon MI, Bi IV, Schroeder SG, Sanchez-Villeda H, Doebley JF et al (2005) A large-scale screen for artificial selection in maize identifies candidate agronomic loci for domestication and crop improvement. Plant Cell 17:2859–2872
143. Tuberosa R, Gill BS, Quarrie SA (2002) Cereal genomics: ushering in a brave new world. Plant Mol Biol 48:445–449
144. Druka A, Druka I, Centeno AG, Li H, Sun Z, Thomas WTB et al (2008) Towards systems genetic analyses in barley: integration of phenotypic, expression and genotype data into GeneNetwork. BMC Genet 9:73
145. Hochholdinger F, Tuberosa R (2009) Genetic and genomic dissection of maize root development and architecture. Curr Opin Plant Biol 12:172–177
146. Urano K, Kurihara Y, Seki M, Shinozaki K (2010) 'Omics' analyses of regulatory networks in plant abiotic stress responses. Curr Opin Plant Biol 13:132–138
147. Ozturk ZN, Talamè V, Deyholos M, Michalowski CB, Galbraith DW, Gozukirmizi N et al (2002) Monitoring large-scale

changes in transcript abundance in drought- and salt-stressed barley. Plant Mol Biol 48:551–573

148. Zinselmeier C, Sun YJ, Helentjaris T, Beatty M, Yang S, Smith H et al (2002) The use of gene expression profiling to dissect the stress sensitivity of reproductive development in maize. Field Crop Res 75:111–121

149. Schnable PS, Hochholdinger F, Nakazono M (2004) Global expression profiling applied to plant development. Curr Opin Plant Biol 7:50–56

150. Talame V, Ozturk NZ, Bohnert HJ, Tuberosa R (2007) Barley transcript profiles under dehydration shock and drought stress treatments: a comparative analysis. J Exp Bot 58:229–240

151. Kofler R, Torres TT, Lelley T, Schlotterer C (2009) PanGEA: identification of allele specific gene expression using the 454 technology. BMC Bioinformatics 10:143

152. Hochholdinger F, Woll K, Guo L, Schnable PS (2005) The accumulation of abundant soluble proteins changes early in the development of the primary roots of maize (*Zea mays* L.). Proteomics 5:4885–4893

153. Fernie AR, Schauer N (2009) Metabolomics-assisted breeding: a viable option for crop improvement? Trends Genet 25:39–48

154. Saito K, Matsuda F (2010) Metabolomics for functional genomics, systems biology, and biotechnology. Annu Rev Plant Biol 61:463–489

155. Sakurai N, Shibata D (2006) KaPPA-view for integrating quantitative transcriptomic and metabolomic data on plant metabolic pathway maps. J Pestic Sci 31:293–295

156. Pelleschi S, Guy S, Kim JY, Pointe C, Mahe A, Barthes L et al (1999) Ivr2, a candidate gene for a QTL of vacuolar invertase activity in maize leaves. Gene-specific expression under water stress. Plant Mol Biol 39:373–380

157. Pelleschi S, Leonardi A, Rocher JP, Cornic G, de Vienne D, Thévenot C et al (2006) Analysis of the relationships between growth, photosynthesis and carbohydrate metabolism using quantitative trait loci (QTLs) in young maize plants subjected to water deprivation. Mol Breeding 17:21–39

158. Boyer JS, McLaughlin JE (2007) Functional reversion to identify controlling genes in multigenic responses: analysis of floral abortion. J Exp Bot 58:267–277

159. Talame V, Bovina R, Sanguineti MC, Tuberosa R, Lundqvist U, Salvi S (2008) TILLMore, a resource for the discovery of chemically induced mutants in barley. Plant Biotechnol J 6:477–485

160. McMullen MD, Kresovich S, Villeda HS, Bradbury P, Li HH, Sun Q et al (2009) Genetic properties of the maize nested association mapping population. Science 325:737–740

161. Zhou M (2011) Accurate phenotyping reveals better QTL for waterlogging tolerance in barley. Plant Breeding 130:203–208

162. Maccaferri M, Sanguineti MC, Corneti S, Ortega JLA, Ben Salem M, Bort J et al (2008) Quantitative trait loci for grain yield and adaptation of durum wheat (*Triticum durum* Desf.) across a wide range of water availability. Genetics 178:489–511

163. Salvi S, Corneti S, Bellotti M, Carraro N, Sanguineti MC, Castelletti S, Tuberosa R (2011) Genetic dissection of maize phenology using an intraspecific introgression library. BMC Plant Biol 11:4

Books and Reviews

Bernardo R (ed) (2010) Breeding for quantitative traits in plants, 2nd edn. Stemma Press, Woodbury, p 400

Costa de Oliveira A, Varshney RK (eds) (2011) Root genomics. Springer, Berlin/Heidelberg, p 318

Guimarães E, Ruane J, Scherf B, Sonnino A, Dargie J (eds) (2007) Marker-assisted selection. current status and future perspectives in crops, livestock, forestry and fish. Food and Agriculture Organization, Rome, p 492

Kole C (ed) (2006) Genome mapping and molecular breeding in plants – cereals and millets. Springer, Berlin/Heidelberg, p 349

Lamkey KR, Lee M (eds) (2006) Plant breeding: the Arnel R. Hallauer international symposium. Blackwell, Ames, p 379

Liu BH (ed) (1998) Statistical genomics: linkage, mapping, and QTL analysis. CRC Press, Boca Raton, p 611

Rao DC, Gu CC (eds) (2008) Genetic dissection of complex traits, 2nd edn. Academic, New York, p 760

Varshney RK, Tuberosa R (eds) (2007) Genomics-assisted crop improvement, volume 1: genomics approaches and platforms. Springer, Dordrecht, p 386

Varshney RK, Tuberosa R (eds) (2007) Genomic-assisted crop improvement, volume 2: genomics applications in crops. Springer, Dordrecht, p 509

Xu Y (ed) (2010) Molecular plant breeding. CABI, Wallingford, p 734

Martin Waste-to-Energy Technology

Johannes J. E. Martin, Ralf Koralewska
Martin GmbH für Umwelt- und Energietechnik,
Munich, Germany

Article Outline

Glossary

Reverse-acting grate Grate system, inclined at an angle of 26°, with rows of grate bars moving up and down against the downward flow of solids.

Horizontal grate Horizontal grate system with rows of grate bars moving in opposite directions alternated with stationary rows of grate bars.

SYNCOM *SYN*thetic *COM*bustion using oxygen-enriched underfire air on a reverse-acting grate.

IR camera Infrared camera recording the surface temperature across the width and the length of the bed on the grate for selected bandwidths from the roof of the combustion chamber.

MICC MARTIN Infrared Combustion Control for reverse-acting grate systems including fuzzy logic control, IR camera, operating mode concept, and operational data logging/visualization.

ACC Advanced Combustion Control for horizontal grate systems combining existing standard measurement information to generate the control loops.

LN Low NO_x technology for significant reduction of NO_x concentration downstream of the combustion system by primary measures.

SNCR Selective Non-Catalytic Reduction of NO_x by injection of a reagent into the combustion chamber as a secondary measure without use of a catalytic converter.

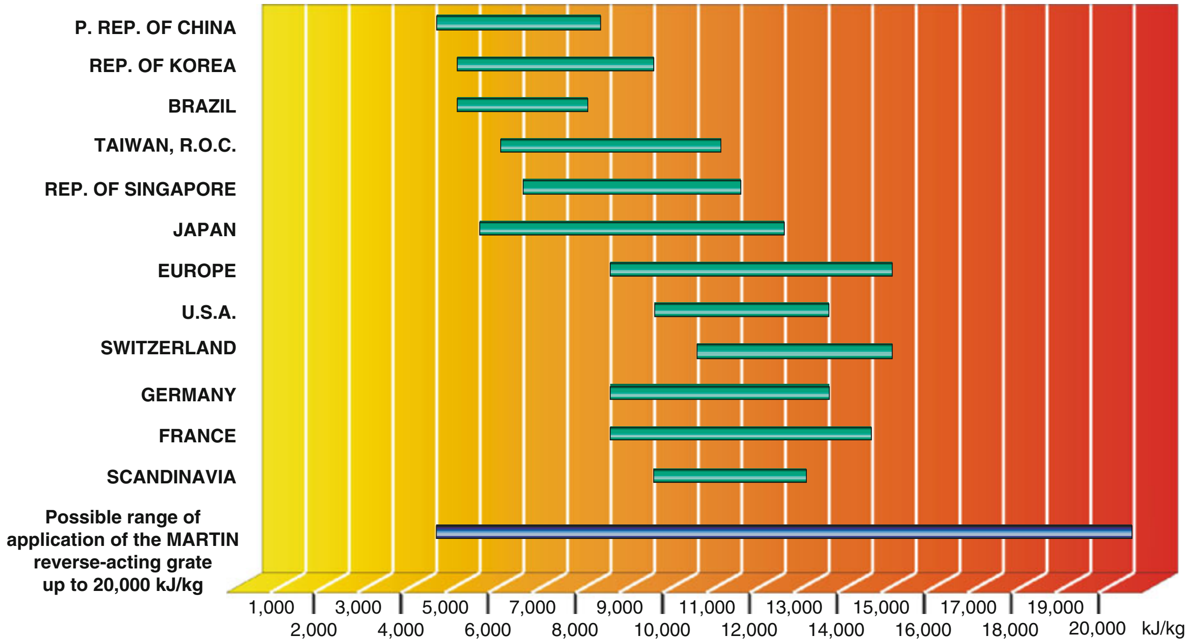
T3 Time–Temperature–Turbulence: concept ensuring an efficient post-combustion with good gas burnout.

Definition of the Subject

Waste-to-Energy (WTE) technologies have been confronted with numerous and changing challenges over the last few decades. Various important factors have to be considered, not only the reduction of waste volume and mass and the destruction and capture of pollutants. Environmental concerns have demanded that flue gases are no longer a significant source of emissions, and that the waste, and also residues remaining after thermal treatment, are to be transformed into reusable products. Thermal treatment of waste using a grate-based system has gained acceptance as the preferred system for sustainable treatment of waste worldwide. The reason for this is that the energy content of the waste is utilized and that quality products and residues are produced. Nevertheless grate-based processes must also keep pace with international requirements and proposed alternative thermal treatment technologies by further innovative development.

Recent years have seen significant increases in the average heating values of waste in many countries (Fig. 1). Essentially, this can be attributed to recycling measures, separate recovery of waste streams, as well as pretreatment and processing procedures in modified waste management concepts. As a result of the above and due to the increasingly frequent application of energy recovery schemes for commercial waste, the fuel input to WTE plants combusting household waste has been significantly influenced. In response to the associated increased thermal load, water-cooling of grate bars, using different technical solutions, was developed and used for forward-acting grate systems.

In view of the continual depletion of raw materials, sustainable processes for the recovery of recyclables are becoming increasingly important. Efficient use of the energy recovered (in the form of electricity, process steam, and district heating/cooling) by developing innovative concepts is also in demand. In Europe, there is potential to increase the contribution of WTE to over 10% of the overall renewable energy produced, since over one half of the energy contained in municipal waste is of biogenic origin. All of these aspects have been driven by political and public pressure, regulations, as well as the financial market situation. Any new demand leads to an increase in the overall complexity of the thermal treatment process. Additional skills and competences are needed for plant design, process control, and operation. Nevertheless, cost-benefit and eco-efficiency analyses clearly show that these additional efforts should be made. Modern WTE plants are extremely complex in terms of the technology used. A sound knowledge of the "fuel" waste as well as its effects on design and operation is crucial for successful planning and implementation of WTE plants.



Martin Waste-to-Energy Technology. Figure 1

Typical lower heating values of untreated household waste in different countries

In addition to the environmentally friendly treatment of waste, its use as a source of energy plays an increasingly important role. As discussions on climate change, diminishing sources of fossil fuels and heavy dependence on their supply from potentially unstable regions intensify, waste is progressively being seen as a resource. On a political level, this has led to greater acceptance of combustion as an indispensable component of sustainable waste management.

Introduction

MARTIN GmbH was established in 1925 in Munich, Germany. The founder, Josef Martin, had already been working on various combustion system designs for coal and waste of all types. The invention of the reverse-acting grate is based on the fact that fuel ignites more easily when an already existing glowing mass is pushed back underneath it. The combined effect of gravity and reverse-acting motion resulted in the fuel being circulated and optimally burned out. The reverse-acting grate and the newly invented ram-type discharger were tested in a pilot unit for the WTE plant in Romainville near Paris in 1932. The combustion tests were successfully carried out using waste from Paris, confirming that the principle on which the grate is based can also be successfully applied to the combustion of municipal waste. In 1952, the company was awarded a contract to build its first WTE plant in São Paolo, Brazil. The grate system proved to be capable of ensuring optimum reaction conditions for combusting municipal solid waste (MSW) and high reliability of the overall operation.

In the beginning of the 1980s, waste combustion technology underwent a fundamental technological change. Step by step, the conventional combustion plant was reengineered to become a modern high-tech power station for treating residual waste. The main forces behind this change were growing environmental awareness on the part of the general public, which was reflected in increasingly strict limit values for emissions; great public interest, which focused the attention of politicians and the media on WTE plants; the constantly changing quality of MSW, which led to increasingly higher heating values; and the demand for efficient materials recycling and energy recovery from residual waste.

However, it soon became clear that the most significant innovations for optimizing combustion on the grate could be achieved through appropriate furnace dimensions and a comprehensive, intelligent combustion control system. Waste is a complex, heterogeneous fuel, the composition and heating value of which fluctuate significantly. Also, the radical change in waste composition in the 1990s, whereby a large fraction of the commercial waste was not recyclable and had a high heat content, necessitated a phase of intense development of existing technologies. Either industrial or biomass waste could be used as fuel.

For over 80 years, MARTIN has been successfully active – as a general contractor, consortium partner, supplier of components, or engineering partner – in the field of combusting difficult fuels such as low-grade coal and waste. Today more than 700 lines for combusting waste have been equipped with MARTIN grates in over 370 plants worldwide [1]. Following restructuring at one of its competitors, MARTIN also took over their horizontal grate technology in 2002.

Components of a WTE Plant

Figure 2 shows the typical components of a WTE plant using the MARTIN technology. A detailed description of the technologies used is provided in the following sections.

In the "tipping" hall, the MSW collection trucks discharge ("tip") their load into the waste bunker, where it is mixed by an overhead crane and is temporarily stored for several days. The tipping hall and waste bunker are totally enclosed buildings to prevent dust and odors from escaping into the environment. In addition, air is continually extracted to maintain a slightly negative pressure and is used as combustion air for the combustion system.

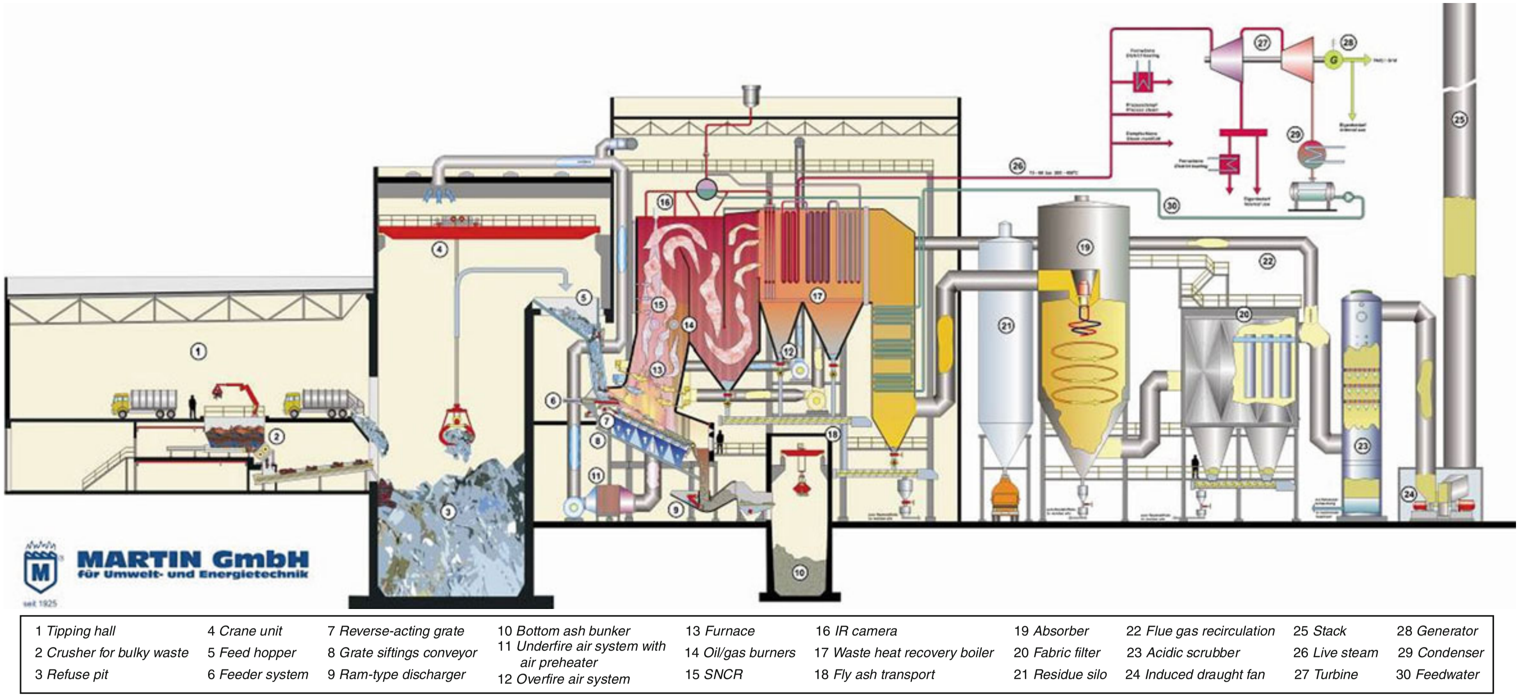
A crane unit lifts the waste from the waste bunker and transports it to the feed system, which consists of a hopper and a feed chute. Feed rams push the waste from the bottom of the feed chute onto the combustion grate. Due to the height of the waste column in the feed chute, unwanted air cannot leak into the combustion system. Microwave level detectors report the height of the waste column in the feed chute to the crane operator. Bridging and obstructions are prevented by the inclined side walls of the hopper and the flaring of the feed chute. A shutoff damper located underneath the hopper is closed when the plant is not operating.

The feed ram changes the direction of waste flow from vertical to horizontal. The waste, compacted in the feed chute, is loosened during this process and pushed onto the grate in amounts determined by the combustion control system. The transition between the feeder and grate can be designed as an inclination or as a drop-off edge. Each feed ram is driven by a hydraulic cylinder. The combustion control system optimizes cycle time, stroke length, and stroke speed to achieve uniform combustion on the grate. The operation of the feed rams is staggered for combustion grates with several parallel grate runs.

The combustion system, consisting of the grate system and furnace, is the heart of the WTE plant. The air required for combustion is fed as "underfire" or primary air from below the grate, passing through the grate bars into the fuel bed. Flue gases emitted from the fuel bed are not completely burned out, therefore "overfire" or secondary air is provided for combustion of the volatile gases rising from the grate. Overfire air is injected into the furnace above the fuel bed via numerous nozzles arranged opposite each other on the front and rear walls of the furnace. The resulting turbulence mixes the flue gas very efficiently, causing complete burnout at temperatures between 1,000°C and 1,200°C. Recirculated flue gas may also be used for secondary combustion. This involves splitting part of the flue gas after the dust removal equipment and returning it to the furnace to replace some of the overfire air. The recirculated flue gas is injected into the furnace via separate nozzles.

The hot, burned-out bottom ash drops from the grate end to a water bath in the ash discharger below the grate where complete quenching occurs. The discharger is filled with water up to the level of the air sealing wall. This creates an air seal against the furnace, thus preventing flue gas and thermal pollution in the basement, on one hand, and air infiltration into the boiler, on the other. The ash discharging ram pushes the bottom ash under the air sealing wall toward the drop-off edge.

Thermal treatment of waste produces energy that is transferred to steam and used in a steam turbine to generate electricity, heat for district heating purposes, or process steam in almost any number of combinations, as described in more detail further on.



Martin Waste-to-Energy Technology. Figure 2
Longitudinal section of a WTE plant

The flue gas cleaning system is an integral and most important part of every modern WTE plant. During combustion, pollutants contained in the waste are released and pass into the flue gas which must be cleaned before it can safely be allowed to enter the atmosphere. Suitable processes and components are combined according to technical and economic requirements. Electrostatic precipitators, fabric filters, spray absorbers, scrubbers, activated coke processes, SCR, or SNCR systems are used in various combinations. The clean gas emissions at the stack are continuously measured, monitored, and documented with state-of-the-art measurement devices. Modern WTE plants respect the strictest emission levels, worldwide, and are trendsetters for the highest environmental standards.

Grate Technologies

The grate system consists of the feed hopper, feeder, combustion grate, bottom ash discharger, and air supply to the furnace (Fig. 3).

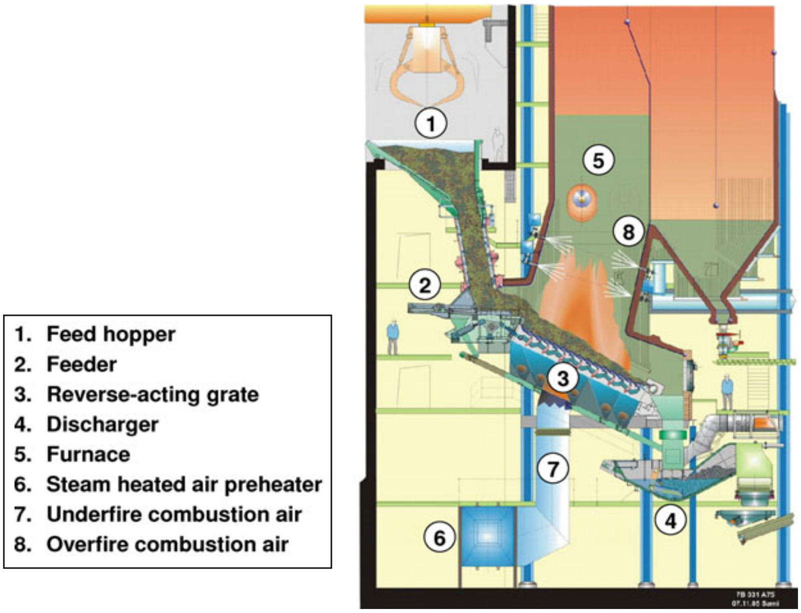
Feeding System

The waste feeding system consists of the charging hopper, connected feed chute with shutoff damper and water cooling in the lower part and ram-type feeders (Fig. 3). The waste taken up by the grapple of the waste crane is fed into the feed chute via the hopper. The dimensions of the hopper ensure that the contents of the grapple can be charged without difficulty and that there is always a sufficient stock of fuel in the hopper. By means of the feed chute, the discontinuous charging of the waste into the hopper is transformed to a continuous flow of fuel. The fine metering is done by the ram-type feeding system arranged downstream of the feed chute.

The design of the feed chute must satisfy the following requirements:

- Operational reliability
- Avoidance of damage and operational failures
- Supply of fuel without blockages

Due to the column of waste it contains, a feed chute of sufficient height prevents the ingress of undesirable air into the furnace and provides a certain stock of fuel in the eventuality of failures in preceding systems. The feed chute is provided with a water jacket. By means of natural convection, the water is moved through the warm and cold zones, thereby preventing overheating. The closed water circuit does not lead to any additional


1. Feed hopper
2. Feeder
3. Reverse-acting grate
4. Discharger
5. Furnace
6. Steam heated air preheater
7. Underfire combustion air
8. Overfire combustion air
1
2
3
4
5
6
7
8

Martin Waste-to-Energy Technology. Figure 3
Typical grate system

water consumption during normal operation. The cooling system for the feed chute is equipped with a thermostat that opens an emergency water connection for cooling, in case the temperature of the water jacket exceeds a certain level, and a warning signal is sent to the control room.

A "bridge" breaking system is provided at the rear side of the hopper to remove any bridging of waste in the transition area between the hopper and the feed chute. Also, in order to be able to take countermeasures in the event of dust formation, an atomized water spray can be generated by means of spray nozzles that are arranged above the upper waste hopper edge. Under the feed hopper, there is a shutoff damper extending over the entire width of the feed chute; this damper is closed when the level in the feed chute drops too low or when the plant is shut down. In order to ensure a sufficiently high waste column in the feed chute, there is a built-in level monitor in the form of microwave sensors.

The direction of movement of the fuel is changed from vertical in the feed chute to horizontal on the feed table. This is done by means of ram-type pushing devices. In the course of charging, a uniform, continuous supply of waste onto the grate is achieved. The design of the combustion grate allows exact fuel metering in accordance with the required thermal output because the feed rams are driven separately for each run via hydraulic cylinders. Cycle time, stroke length, and stroke speed can be varied, independently for each run. The waste transfer zone at the end of the feeding table is designed with a gentle drop-off edge, which has the effect that the fuel is loosened with a minimum of dust formation.

The feeding system is part of the combustion control concept; however, it can also be controlled manually and the feeding ram can be switched from "working" stroke to "clearing" stroke, enabling the feeding system to be emptied within the shortest possible time.

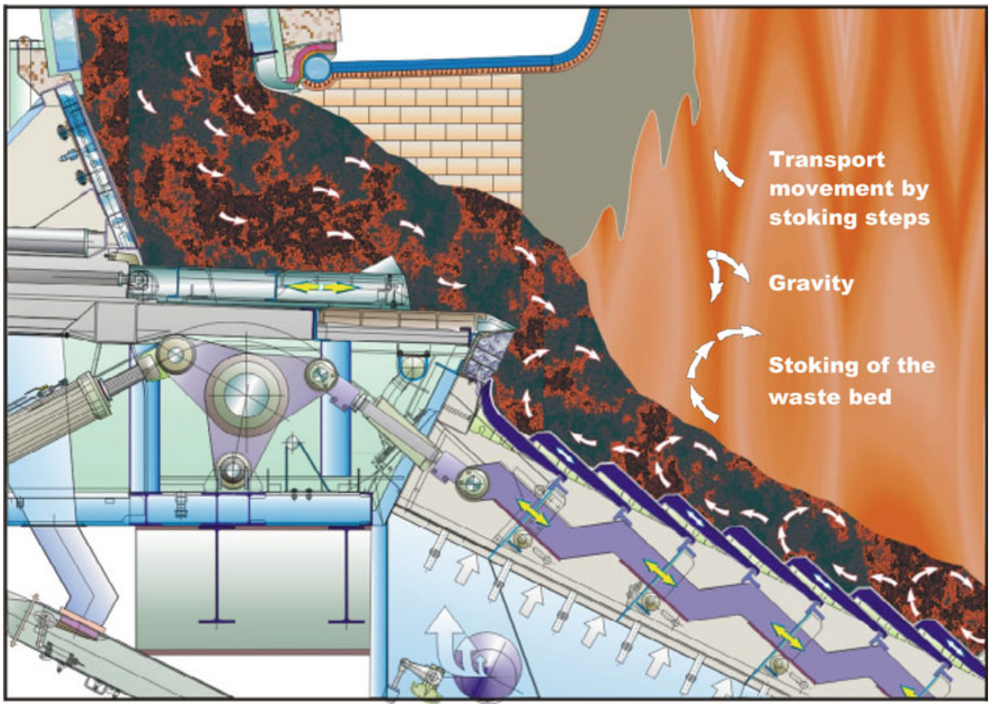
Reverse-Acting Grate

The reverse-acting grate is inclined at an angle of 26° to the horizontal and consists of several stair-like grate "steps" that consist of bars. Every second step is moved up and down against the downward direction of the solids flow (Fig. 4). The moving rows of grate bars are moved forward and backward by a hydraulic cylinder. This constantly rakes and mixes the red hot mass with newly fed waste. In the area of the feeding system, predrying of waste is carried out by radiation from the flame above the grate. The drying process is fully completed in the front grate area. The waste begins to burn already at the front end of the grate and the fuel bed temperature reaches 1,000°C and higher. The waste is combusted fully during its travel over the length of the grate. The combustible constituents of the fuel are converted to gases in the main combustion zone accompanied by the release of energy (primary combustion). Secondary combustion, e.g., the final oxidation of the unburned gases, takes place in the flame above the main combustion zone. Complete gas burnout is achieved by means of overfire air injection that is optimally controlled according to the prevailing furnace conditions.

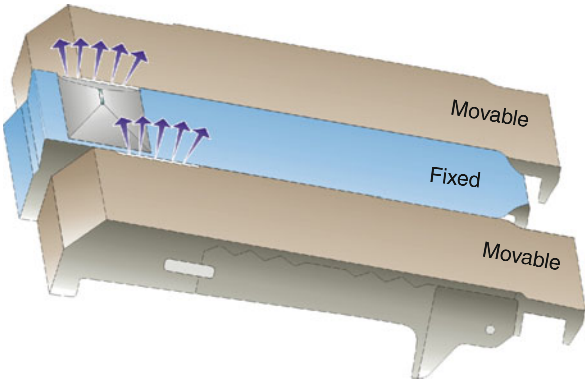
The number of these agitation cycles, in other words the grate speed, is primarily dependent on the composition of the fuel and only to a minor extent on the combustion throughput. The residence time of the fuel on the grate is typically between 60 and 70 min.

Along its length, the reverse-acting grate is divided into three to six separate air zones, so that underfire air is supplied across the grate in a controlled manner and as needed for combustion. The underfire air flows into the fuel bed through narrow gaps at the head of the grate bars (Fig. 5). These air gaps are kept free of impurities during operation because every second grate bar in a row moves, relatively to its adjacent bars, at the end of each agitation stroke. The effect of the relative stroke is to clear the air gaps, thereby allowing a long period between maintenance outages. The evenly spaced narrow air gaps ensure that the underfire air is distributed evenly over the fuel bed. The grate bars are made of cast chromium steel that is highly resistant to wear.

Combustion on the grate is completed after approximately two thirds of the length of the grate. A clear delimitation between the combustion zone and the end zone is plainly visible. Through its intimate mixing of the fuel, the reverse-acting grate always ensures good thermal protection for the grate bars due to the "insulating" fuel and ash layer on the grate surface. The average operating temperature of the grate bars is approximately 20–50°C above the underfire air



Martin Waste-to-Energy Technology. Figure 4
Mixing of the waste on a reverse-acting grate



Martin Waste-to-Energy Technology. Figure 5
Underfire air flow/movement of grate bars of a reverse-acting grate

temperature even in the main combustion zone of the grate. This provides a high level of certainty against thermal overloading and makes it possible to achieve a long service life for the grate. The operating experience from numerous plants obtained over many years of continuous service time have conclusively proven that the reverse-acting grate does not require water cooling, even for high heating-value fuel (Fig. 1) [2].

The grate "siftings," i.e., inert particles that may fall through air gaps between the grate bars (mostly in the residue burnout area) are collected in the hoppers under the grate. The particles are discharged from the hoppers by means of pneumatically operated shutoff valves, with no air infiltration. The shutoff valves operate according to a specified timing program to open the ducts and discharge siftings into the bottom ash discharger where they mix with the bottom ash.

The reverse-acting grate is of modular design (Fig. 6). Each module comprises a complete grate run with a width of 1.5–2.5 m. The modules can be fully preassembled at the factory and then shipped to the plant site. Up to eight grate-run modules can be arranged in parallel to produce a total grate width of over 15 m.

A clinker roller, or clinker weir, is installed at the end of the grate to control the height of the fuel bed and bottom ash layer. The roller, or weir, can be adjusted to suit actual combustion conditions. From there, the bottom ash drops into the bottom ash discharger.

Martin Waste-to-Energy Technology. Figure 6
Reverse-acting grate

The reverse-acting grate satisfies the following essential requirements of MSW combustion:

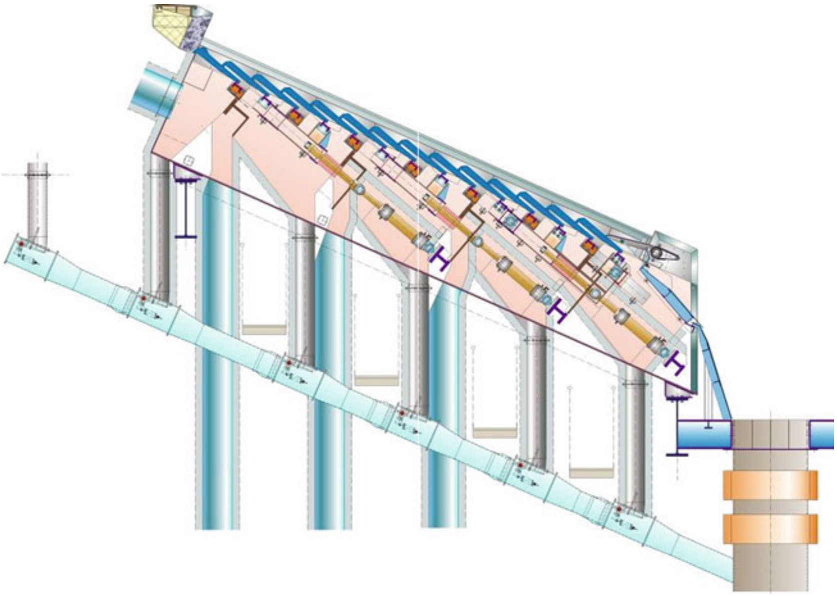
- Wide heating value range and capacity for handling fluctuations in waste composition
- Rapid transition of the fuel from a cold state into the combustion phase in order to prevent smoldering that adversely affects emissions
- High and uniform fuel bed temperature
- Intensive and constant agitation of the fuel bed
- Clear demarcation between combustion zone and burned-out bottom ash over the entire grate width
- Uniform coverage of the grate surface
- Easily definable and controllable supply of underfire air to suit requirements
- No air supply due to mechanical requirements (e.g., cooling of cast-steel parts)
- Low thermal load on grate surface due to covering of the bars with bottom ash
- Small amount of grate siftings
- Low dust emission from the combustion process
- Direct response to control operations
- High equipment availability
- Rapid start-up and shutdown of the grate
- Easy replacement of grate bars

The overall result is long service life for the grate, high availability, ability to recover residues, and compliance with emission requirements.

Reverse-Acting Grate Vario

The reverse-acting grate Vario (Fig. 7) was developed for use with fuels with a high heating value and a low ash content (e.g., a refuse-derived fuel, RDF). This grate uses the same proven and unique reverse-acting principle but its angle of inclination is 24°. The stair-like grate steps are alternately arranged in stationary and moving grate bar rows. The interaction between the upward stoking force and the downward pull of gravity ensures constant mixing of the red hot mass with the fresh fuel. This results in optimal combustion and fully burned-out bottom ash. At the same time, the grate is automatically covered with a thick fuel and ash layer which insulates the grate surface and provides excellent protection against thermal radiation from the furnace (>1,100°C).

The reverse-acting grate Vario is divided into three independent drive zones along its length so that full advantage can be taken of the reverse-acting principle



Martin Waste-to-Energy Technology. Figure 7
Reverse-acting grate Vario

even with fuel that has a low inert content. The width of the grate consists of a number of modules, depending on the design capacity of the unit. Each module is driven independently.

The grate Vario satisfies the same requirements as listed earlier for reverse-acting grate. Additionally the following essential requirements are met:

- Grate speed can be individually set per run and zone
- Clear delimitation between combustion zone and burned-out bottom ash over the entire grate width
- Good bottom ash burnout by adjusting the grate speed in the various zones in response to varying waste quality
- Easily controlled supply of combustion air as required
- Small amounts of grate siftings
- Modular design

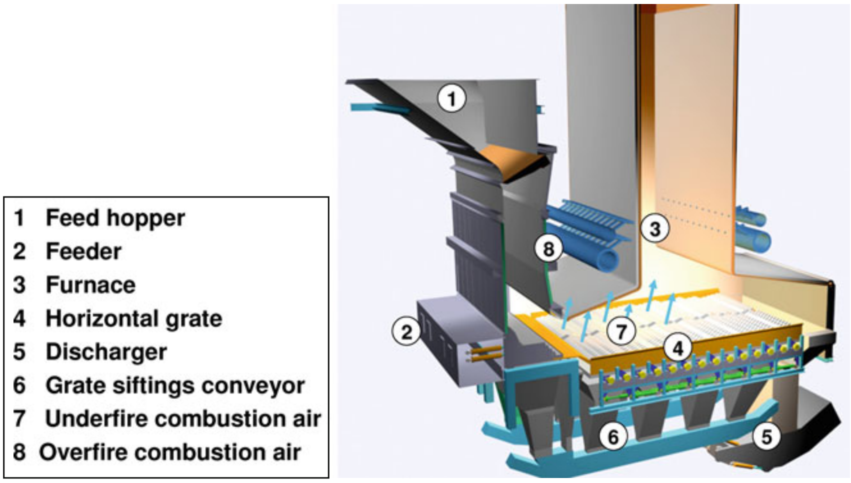
This design results in the following advantages for combustion system operation: high level of availability, long grate surface service life, recyclable residues, and consistent compliance with emission requirements while at the same time maintaining high efficiency.

Horizontal Grate

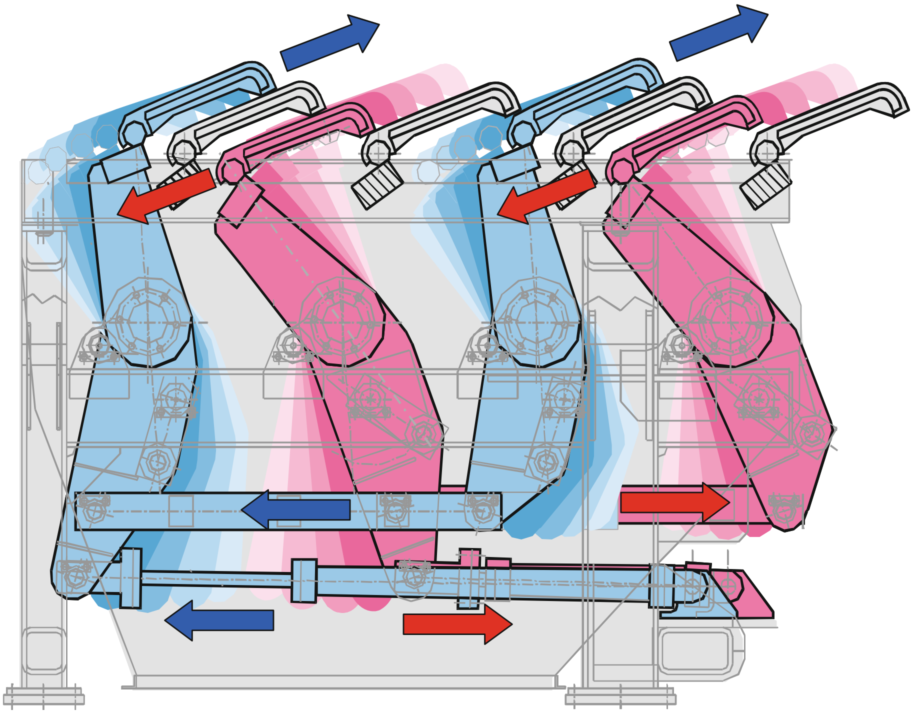
The horizontal grate system was originally developed by the Swiss company W + E Umwelttechnik AG. It has been part of the MARTIN technology portfolio since 2002 (Fig. 8). It was developed for the incineration of household waste and combustible industrial residues. The horizontal construction of the combustion grate allows the fuel to be advanced in a well-controlled manner. The grate is modular, providing for the configuration of a variety of different grate sizes, depending on given variables such as:

- Quantity of waste
- Lower heating value of waste
- Type of use (heat evaluation)
- Structural constraints

The waste is pushed forward by rows of grate bars moving in opposite directions, alternated with stationary rows of grate bars (Fig. 9). Moving and stationary horizontally arranged grate-bar supports are located in the solidly built grate support structure. The moving-bar rows are connected to an oscillating crank and move back and forth, working in opposite directions


1 Feed hopper
2 Feeder
3 Furnace
4 Horizontal grate
5 Discharger
6 Grate siftings conveyor
7 Underfire combustion air
8 Overfire combustion air
1
2
3
4
5
6
7
8

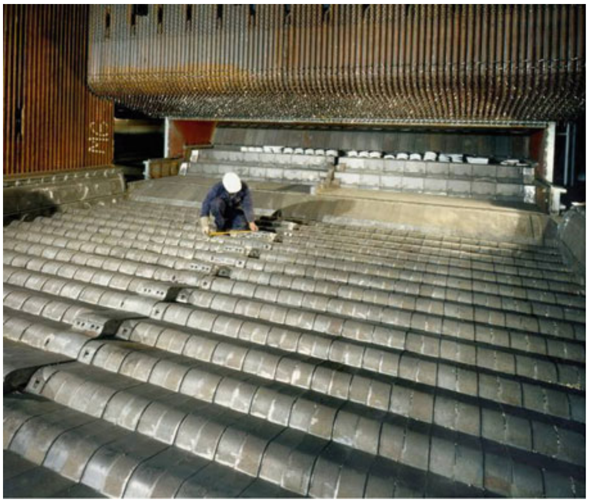
Martin Waste-to-Energy Technology. Figure 8
Horizontal grate system



Martin Waste-to-Energy Technology. Figure 9
Movement of grate bars of a horizontal grate

relative to each other. When the rows of grate bars are moved apart in opposite directions, the fuel is dropped down at this point and ignited fuel particles fall down. The grate bars then move toward each other, lifting and igniting the fuel layer. The fuel particles initially ignited as a result of the backward motion of the bars then move forward as a result of the advance feed motion of the bars and drop to the bottom of the fuel layer, where they become ignition points and accelerate combustion of the fuel above them. On the horizontal grate, ignition takes place not only by radiation and convection of heat from above, but also from within the waste bed. Burning particles are constantly being pushed downward by the intensive stoking and agitation of the waste layer, causing ignition to start from within the bed as well.

The continuous stoking and agitation of the waste results in the waste layer being repeatedly broken up and rearranged. This promotes ignition and combustion by creating a large waste surface area, which favors the admission of combustion air, and effectively transports and mixes the waste to ensure good burnout.



Martin Waste-to-Energy Technology. Figure 10 Horizontal grate

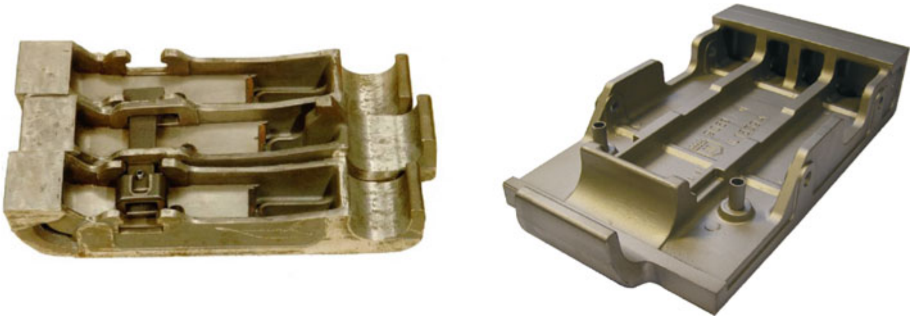
The fuel residence time on the grate (drying, main combustion, and burnout zones) depends on combustion capacity, fuel composition, processes, etc., and takes about 30–120 min. From the grate rear end, bottom ash falls into the water bath of the bottom ash discharger.

The main features of the horizontal grate are:

- Combination of fixed and moving rows of grate bars
- Movement of the grate-bar rows in opposite directions
- Infinitely variable hydraulic drive system
- Slow, continuous movement
- High pressure loss over the grate bars with resultant optimal air distribution
- Low overall height
- No grate steps
- No maintenance of the underfire air area

The horizontal grate is of modular design. The length of each module is fixed but the width may vary according to specific requirements. Each module has its own drive and supply of underfire air, both of which can be controlled separately. A typical grate configuration consists of three modules in the direction of waste flow and there may be between one to three runs in parallel depending on the waste throughput capacity (Fig. 10).

The grate bars are typically air-cooled (Fig. 11) by the flow of underfire air passing through a labyrinth of passages, cooling the bar and promoting a uniform release of heat. The underfire air exits from the front end of the grate bar. The outlet slots have been designed and positioned for an optimal supply of air to the fuel



Martin Waste-to-Energy Technology. Figure 11 Air-cooled and water-cooled grate bars bed. During operation, the air outlet slots are kept clear by the relative motion of adjacent grate bars. The grate bars are made from wear and heat-resistant chromium-nickel alloy steel. The grate bars are an exact fit, which means that the proportion of grate siftings to waste is low. The bottom ash which drops through the grate is collected in ash hoppers and discharged into the bottom ash discharger.

In recent years, there have been significant increases in the heating values of wastes, necessitating the development of water-cooled grate bars (Fig. 11) to achieve acceptable service lives. The solution is a cast-steel construction, where one water-cooled grate-bar block has the width of three air-cooled ones. The cooling circuit is a closed-loop cooling circuit. Two different versions have been installed:

- Medium pressure version (around 7 bar(a)), where the heat removed is then used to either preheat the combustion air or the condensate
- High-pressure version, where the cooling circuit is part of the waste heat boiler circuit; the heat removed is fed to the water–steam cycle

An added benefit of the water-cooled system is that underfire air is no longer needed to cool the grate bars and is controlled only as and when required by the combustion process. The heat dissipated by means of the water-cooling system can be returned in full to the process.

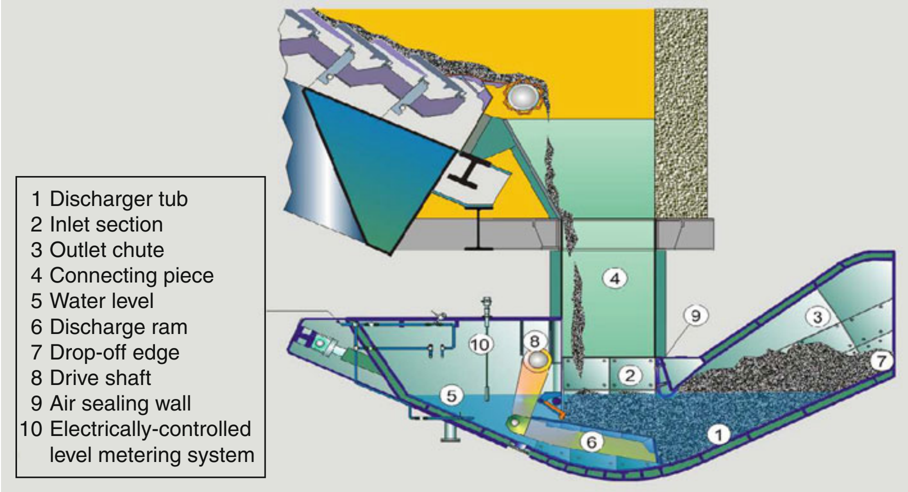
The horizontal grate is notable for a number of beneficial characteristics:

- Controlled transport of the waste due to double-motion overthrust mechanism of the grate bars
- Horizontal construction allows the waste to travel in a controlled manner and precludes sliding of the waste
- Tight-fitting grate bars with high pressure drop producing an even intake distribution of the underfire air over the width of the grate
- Small percentage of combustion products ending up as siftings
- Parts subject to wear are made of a high-tensile cast alloy steel
- A high level of availability and operational safety
- Easy maintenance

Bottom Ash Discharger

From the rear end of the grate, the bottom ash falls into the water bath of the bottom ash discharger (Fig. 12).

The discharging ram pushes the bottom ash under the air sealing wall toward the drop-off edge. As a result, the bottom ash is discharged in a dust-free and odorless manner. Larger pieces of bottom ash



Martin Waste-to-Energy Technology. Figure 12
Bottom ash discharger

become cracked as a result of quenching in the water. They are then broken by the force of the discharging ram.

Because the bottom ash still resides for a while in the chute above the actual water level and during that time is compressed by the discharging ram, most of the water flows back into the discharger tub, with the result that the bottom ash is discharged with only low moisture content. The discharger has no water overflow, so only the small amounts of water escaping through evaporation and taken up by the bottom ash need to be replaced. As necessary, water is added and controlled via a float valve. In normal operation, therefore, the discharger has no water drain and has a very low water consumption. In this way, clean operation is ensured, while at the same time a very sturdy, space-saving, and easily accessible design is obtained.

The interior of the discharger tub and the side walls of the inlet chute are lined with wear plates. The front plate of the discharging ram is protected by easily replaceable slide strips.

A wide range of dischargers is available, the smallest with a discharging capacity of approximately 0.2 m^3/h. Dischargers for coal-based combustion systems have a discharging capacity of approximately 0.2–3.0 m^3/h. Discharging capacities for waste combustion plants range between approximately 4.5 and 12.0 m^3/h.

Combustion Technologies

Using optimal combustion design in terms of thermal efficiency and temperature levels, systems must be applied to limit the specific global and local heat loads to reasonable values for system components, i.e., grate, combustion chamber, and post-combustion chamber, taking design criteria into account:

- Optimal mixing of the flue gases
- Homogeneous temperature and concentration profiles
- Stable combustion and constant heat release
- Minimum formation of pollutants
- Reduction of chemical attack and thermal stress
- Reduction of corrosion and abrasion of materials
- Optimal burnout of bottom ash

A combustion system capable of maintaining stable and uniform operation in terms of thermal load, flue gas flow, bottom ash burnout, and flue gas burnout has major advantages compared with systems in which conditions fluctuate:

- Easier compliance with regulations and limit values
- Smaller design range required for flue gas cleaning components
- Minimum operational effort required for combustion system control
- More uniform thermal and mechanical load and, consequently, longer service life for components such as feeder, grate bars, etc.

Conventional Combustion with Air

The conventional combustion air system essentially comprises the underfire and overfire air systems. Optimum interaction between these systems enables combustion to be as homogeneous as possible with low pollutant emissions.

The underfire air is drawn from the building housing the waste bunker. This is done by means of a frequency-controlled fan through which the underfire air flow rate is adapted to the prevailing requirements of combustion on the grate. The extraction of air from the waste bunker has the positive side effect that a slight negative pressure is maintained in the bunker area, causing fresh air to flow in. This ensures that no odor nuisance occurs outside the plant. After optional preheating, the air flows through the grate and in this way is supplied to the fuel bed in accordance with requirements. Preheating is achieved using a steam-heated air preheater with an air temperature depending on the heating value of the waste and the kind of waste to be burned. Normally for air temperatures up to 120°C, 5 bars process steam is used. Higher air temperatures up to 160°C (e.g., at lower heating values and co-combustion of sewage sludge) are achieved by the second staged supplied with saturated steam from the boiler drum.

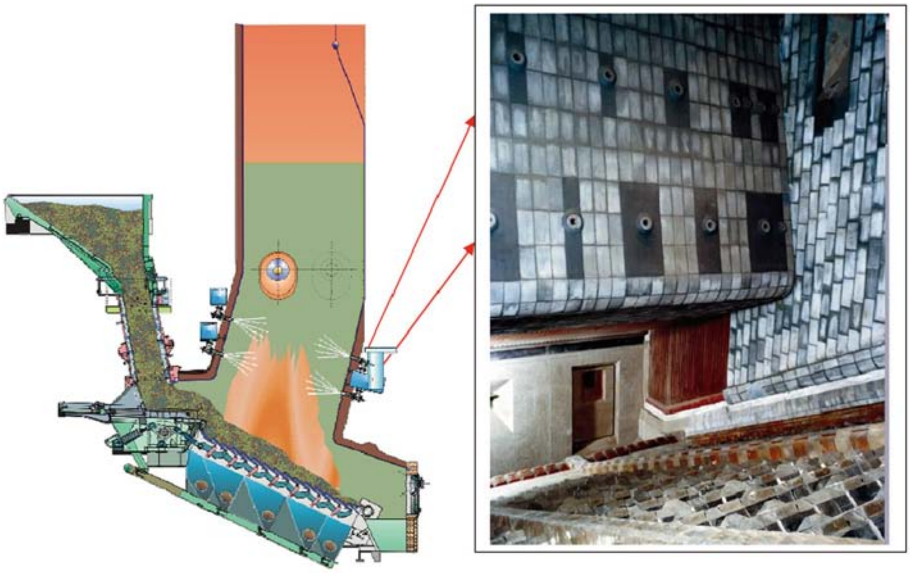
The grate is divided lengthwise into five separate air zones. Air is supplied to the individual zones via orifice openings that are covered by adjustable dampers. The shape and size of the orifice openings are determined according to the combustion characteristics and the designed release of heat over the length of the grate. Accordingly, the largest free orifice opening is located under the grate compartments of the main combustion zone. This concept for underfire air distribution makes it possible to distribute the combustion air to the fuel bed in an individual manner according to requirements. As a result, it is possible to react to differences in the waste quality during combustion.

A vigorous, stable fire, in which all the combustion phases (drying, gasification, ignition, and combustion) occur simultaneously and consecutively, develops at the front end of the grate. The constant stoking motion provides for a uniform heat release and ensures excellent burnout values.

The overfire air is drawn from below the boiler house roof or, alternatively, is extracted from the bottom ash bunker and is injected according to the "stitching method." The overfire air is injected through four rows of nozzles, two rows in the front and two rows in the rear wall of the furnace (Fig. 13). The nozzles in the two rows on each wall are arranged in an offset pattern (stitching). This so-called 4-row stitching arrangement ensures full coverage of the furnace cross section and the turbulence needed for good mixing of the gases, while uniform profiles for temperature and flow are obtained above the injection levels. Gases arising from the fuel bed on the grate are oxidized immediately afterward in the downstream furnace by final mixing of the overfire air with the remaining underfire air at high temperatures. The overfire air is supplied by a frequency-controlled fan, through which the flow rate can be optimally adapted to the prevailing requirements in the furnace.

Conventional Combustion with Recirculation of Flue Gas

Burnout in the fuel bed takes place due to the constant ignition and the supply of underfire air. Most of the unburned gases produced on the grate are oxidized immediately afterward in the furnace after being mixed with the remaining underfire air at high temperatures. However, some unburned gases inevitably escape primary combustion. For this reason, adequate mixing of the combustion gases must be ensured by introducing an additional jet of overfire gas via nozzles. As described above, the "4-row stitching" process for the arrangement of the overfire air nozzles and the recirculated gas nozzles is used (Fig. 13). As a result of this measure, a uniform temperature and flow profile, and optimal mixing of the gases in the furnace is achieved. The residence time of the gases in the high temperature zone is extended and gas burnout improves.



Martin Waste-to-Energy Technology. Figure 13
Positioning of overfire air and/or recirculation flue gas nozzles

In order to reduce the flue gas flow (approximately 20–30%), the excess combustion air and the NO_x content, part of the flue gas flow downstream of the effective pre-deduster (e.g., electrostatic precipitator, baghouse) is extracted with a separate frequency-controlled fan and returned to the furnace as recirculated flue gas. The recirculated flue gas replaces most of the required overfire air. It is injected into the furnace via both rows of nozzles on the rear wall as well as the upper row of nozzles on the front wall.

Since, as described above, unburned gases escape from the fuel, overfire air is injected into the furnace via the lower row of nozzles on the front wall. Sufficient oxygen for the combustion of these gases is provided by the overfire air at this point. The turbulent effect of the overfire air is mainly produced by the recirculated flue gas. The required overfire air could be extracted from the underfire air flow downstream of the air preheater. If the recirculated flue gas is extracted after a baghouse filter, the gas temperature is quite low (about 150°C). In this case, the gas ducts have an electric trace heating system to prevent condensation on the walls. Generally the flue gas ducts are insulated to keep heat loss to a minimum and to protect the outside environment. Since the process requires that the flue gas recirculation system is operated with flue gas, which still contains dust remnants, despite prior dust removal in the pre-deduster (dust content $> 20\ mg/Nm^3$), discontinuous cleaning of the recirculated flue gas nozzles is possible. A fine water jet is injected directly into the nozzles in regular adjustable cycles for short periods, so as to remove any flue gas constituents that may still be present. This system ensures proper functioning of the recirculated flue gas nozzles without having any measurable influences on the operation of the plant (furnace temperature, emissions, etc.).

Combustion with Oxygen-Enriched Air (SYNCOM)

Discussion about how to minimize the level of pollutants produced by WTE plants all over the world has led to many new technological developments, focusing in particular on the treatment of flue gas and residues. Above all, the dioxin input introduced with the waste of about 50 μg TEQ/Mg must be destroyed. Consequently, reduction of pollution burdens, improvement of bottom ash quality and reduction of overall dioxin output to <5 μg TEQ/Mg of waste have been the driving forces behind the development of the SYNCOM process (*SYN*thetic *COM*bustion) [3].

The SYNCOM process is based on the following components, illustrated in Fig. 14:

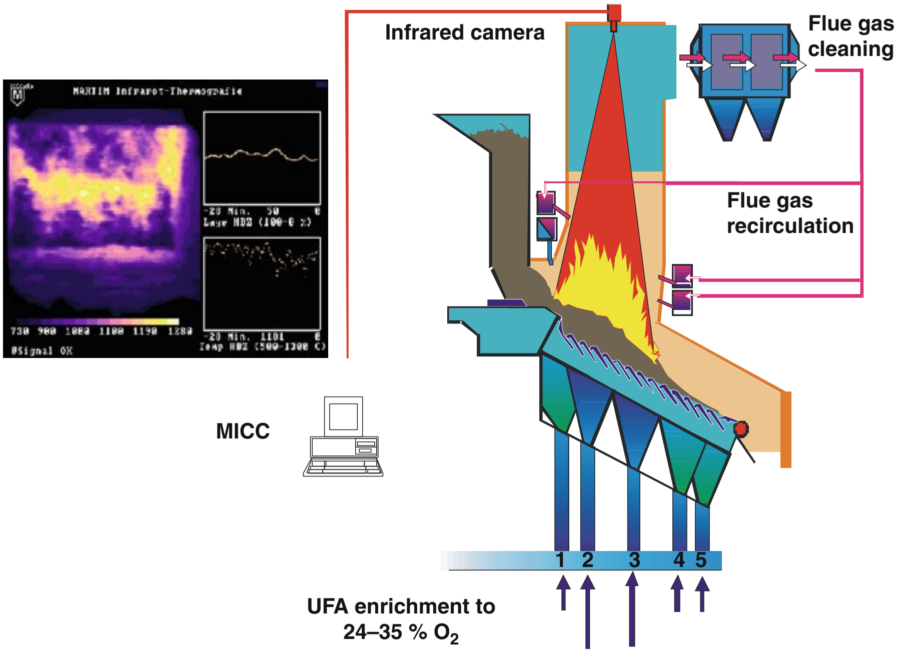
- Grate-based combustion system using the reverse-acting grate
- Combustion control system using IR thermography
- Overfire air system with four nozzle rows (four-row stitching)
- Flue gas recirculation
- Oxygen enrichment of underfire air

The amount of nitrogen in the combustion air is significantly reduced by replacing part of the underfire air in the main combustion zones 2 and 3 with technically pure oxygen. The oxygen concentration in these zones is then in the range 24–35%. As a consequence, the excess air rate is significantly lower, the flue gas flow is substantially reduced and the pollution burden is clearly decreased compared with conventional combustion. The IR-camera control signal is used to meter oxygen selectively and as required.

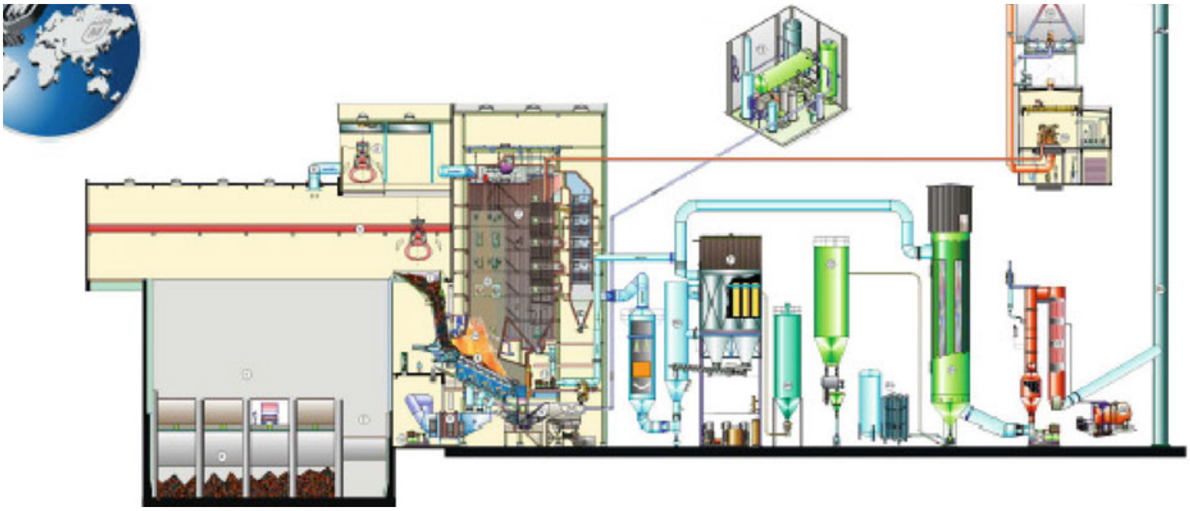
One of the core components of the SYNCOM process is the oxygen generation plant. The oxygen is obtained from the ambient air in an air separation plant. In Arnoldstein (Fig. 15), the air separation plant uses the pressure swing adsorption principle, whereby oxygen and nitrogen are separated by means of a zeolitic molecular sieve. The nitrogen is adsorbed on the molecular sieve, while the oxygen flows through the adsorber. The adsorbed nitrogen, water, and CO_2 are then desorbed by reducing the pressure within the adsorber unit.

The surface temperature of the fuel bed is determined using the IR camera in the boiler roof. The temperature is about 100°C higher than when using underfire air without oxygen enrichment. The higher fuel bed temperature means that bottom ash sintering, improved burnout, and destruction of the organic pollutants are achieved.

To reduce the flue gas flow and O_2/NO_x content, the recirculated flue gas is drawn off downstream of the pre-deduster with a separate fan and returned to the furnace. The recirculated flue gas causes turbulence and mixing. Recirculated flue gas is directed to both nozzle rows on the rear wall and to the upper row of the



Martin Waste-to-Energy Technology. Figure 14
SYNCOM process



Technical data	
Annual throughput approx.:	80,000 Mg
Number of lines:	1
Waste capacity per line:	10.7 Mg/h
Thermal capacity per line:	29.6 MW
Steam output per line:	35.2 Mg/h
Steam pressure:	40 bar
Steam temperature:	400 °C

1 Tipping hall
2 Bulky waste cutter
3 Waste bunker
4 Waste crane
5 Waste crane set down
6 Underfire air intake
7 Feed hopper
8 Reverse-acting grate
9 Discharger
10 Underfire air fan
11 Underfire air preheater
12 Oxygen distribution
13 Air separation plant
14 Overfire air nozzles
15 Flue gas recirculation fan
16 Burners
17 Steam boiler
18 Turbine / generator set
19 Air-cooled condenser
20 Turbo reactor
21 Fabric filter
22 Additive silo
23 Spent active coke silo
24 Nitrogen station
25 Active coke filter
26 ID fan
27 SCR catalytic converter
28 Aqueous ammonia storage
29 Stack

Martin Waste-to-Energy Technology. Figure 15
SYNCOM plant Arnoldstein (AT)

front wall. Due to the use of recirculated flue gas in the overfire air system and the reduction of nitrogen in the underfire air, the gas volume per ton of feed is reduced by up to 35%, as compared with conventional waste combustion.

The smaller flue gas flow has a direct effect on the equipment downstream of the combustion system. As a result, the steam boiler (with the exception of the first boiler pass), any downstream dedusting filter, the ID fan, and the downstream flue gas cleaning equipment can be designed with smaller dimensions than is the case in conventional combustion systems, thereby resulting in lower investment costs and higher energy efficiency. The smaller flue gas volume associated with the SYNCOM process also means that flue gas heat loss is also reduced. The thermal efficiency of the boiler is increased by 3–5%, as compared with conventional thermal waste treatment.

The use of the SYNCOM process results in reducing both capital and operating costs. The additional costs associated with constructing and operating the air separation plant are largely offset by the savings made due to smaller components, a smaller construction volume, better energy utilization, and lower operating costs for the equipment downstream of the combustion system.

In conclusion, the SYNCOM process is characterized by the following features, as compared with conventional combustion using air:

- Intense, uniform combustion
- Minimal CO content in the flue gas
- Temperature in the fuel bed in the main combustion zone approximately 100°C higher
- Partial sintering of the bottom ash and consequently
  - Optimum burnout
  - Minimal heavy-metal leaching in compliance with drinking water quality standards
- Reduction of flue gas flow by approximately 35%
  - Higher boiler efficiency
  - Reduced pollutant burden at stack
  - Reduced fly ash flow

This process was tested extensively in industrial-scale demonstration plants at Coburg (DE), Oita (JP), and Osaka (JP). It was then included in the new WTE plant in Arnoldstein (AT) in 2004 (Fig. 15) and also at the Sendai (JP) WTE plant (three lines).

The WTE plant in Arnoldstein consists of a combustion system with reverse-acting grate (one line, two runs), air separation plant, four-pass vertical boiler, fluidized bed adsorption reactor, fabric filter, lignite-coal fixed-bed filter, ID fan, SCR-DeNO$_x$ unit, stack, turbine, and generator.

Gasification with Post-combustion

Gasification of municipal solid waste has been developed using various technologies, the most widespread being:

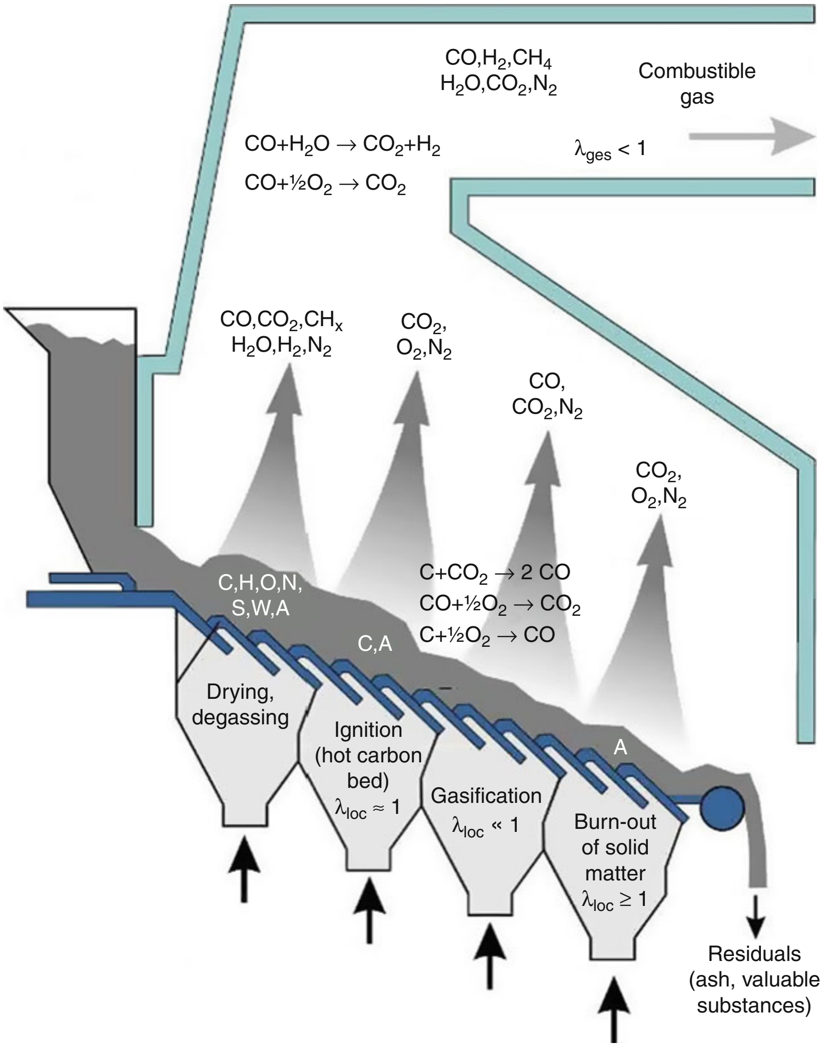
- Shaft furnaces (using additional coke)
- Pyrolysis kilns
- Fluidized bed systems

Experience in Europe in the 1990s showed that these processes have their limitations when used for waste due to high cost, low availability, poor energy efficiency, and the need for waste pretreatment [4–7]. Experience has also proven unstable operation under commercial operating conditions. This is why many European waste gasification plants were dismantled after a relatively short period of time. In the 2000s, a large number of gasification systems were built in Japan and continue to operate despite these limitations [8, 9]. All these systems are coupled with post-combustion of the produced gases in furnaces with boilers. The expected advantage of these gasification systems in Japan is the integrated melting of the ash in the post-combustion step, which also leads to reduction of the total dioxin output (through melting of fly ash) [10, 11].

MARTIN has developed a grate-based gasification system since the 1990s. This gasification system has the advantage of controlling the following conversion steps (Fig. 16):

- Drying
- Degassing
- Ignition
- Gasification
- Burnout of carbon

Based on numerous tests with a semi-industrial-scale combustion unit, it has been proven that the reverse-acting grate (Fig. 3) is essentially well suited to gasification. In particular, the reverse-acting grate



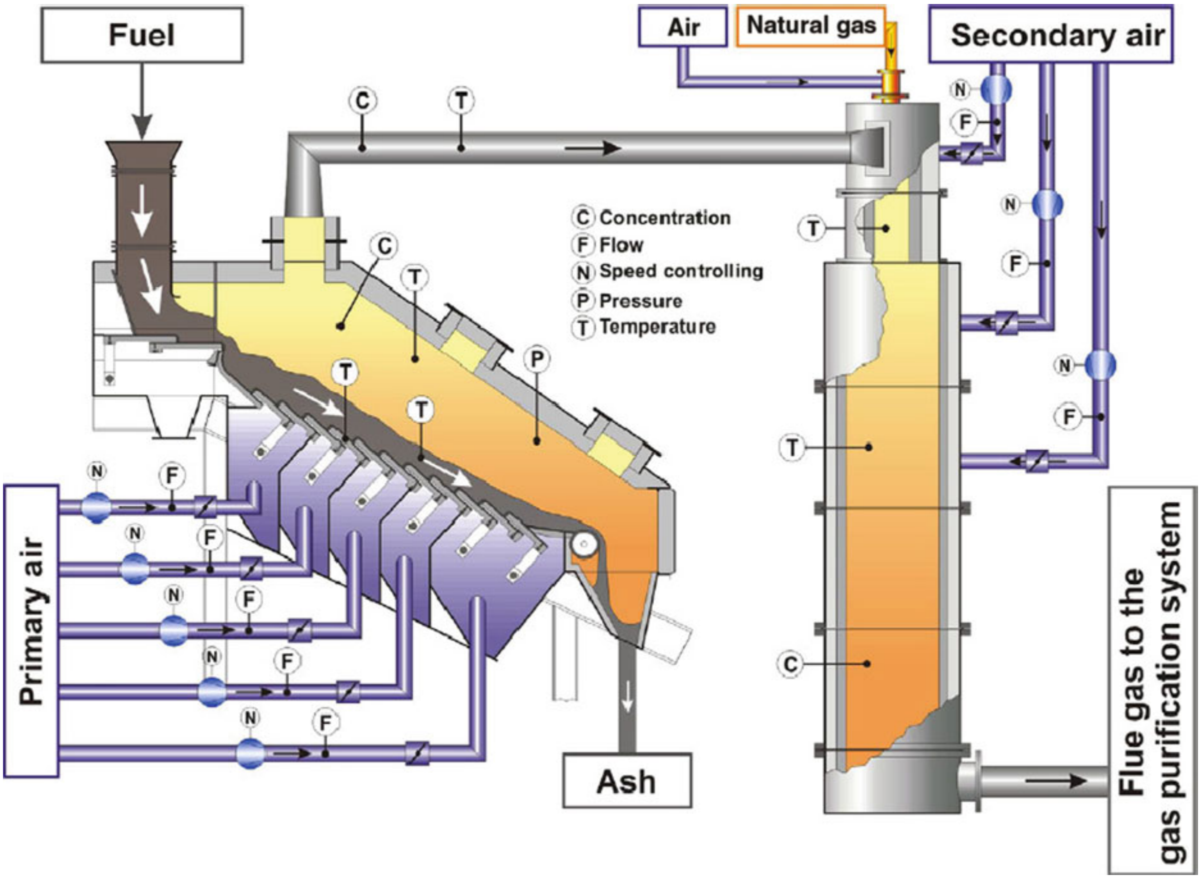
Martin Waste-to-Energy Technology. Figure 16
Conversion steps of grate-based gasification [12]

Vario (Fig. 7) performs well regarding gasification because it has the particular feature that three grate zones can be driven separately and the air compartments are perfectly tight in order to control these conversion steps. Another important advantage of grate-based gasification compared with other gasification systems is its robustness, which makes it possible to use waste that has not been pretreated.

Post-combustion of the gas product of volatilization occurs either in an extension of the membrane wall furnace, or in a following, uncooled separate combustion chamber. Experience with post-combustion chambers is based on the Clausthal test facility (Fig. 17) as well as on the WTE plants in Trieste and Cagliari using T3 (Time–Temperature–Turbulence) controls and Vortex post-combustion chambers (Fig. 18).

Grate-based gasification systems with post-combustion processes are characterized by:

- Reduced excess air rate from 1.8 to 1.4 and thus reduced flue gas flow
- High residence time for sub-stoichiometric flue gas and thus reduced NO_x



Martin Waste-to-Energy Technology. Figure 17
Clausthal test plant for gasification with post-combustion [13]

- Efficient post-combustion (time, temperature, turbulence), resulting in good gas burnout and reduced dioxin content
- Low gas velocities through waste bed and in the gasification chamber; thus reduced fly ash carryover to the boiler
- Taking advantage of the long experience of traditional grate-based systems: moderate cost, high availability, high energy efficiency, and the possibility of treating waste that has not been pretreated

However, the gasification process requires higher waste heating values for stable reaction conditions on the grate and more sophisticated process control, because changes in waste quality have a much greater influence on the process than in the case for superstoichiometric combustion.

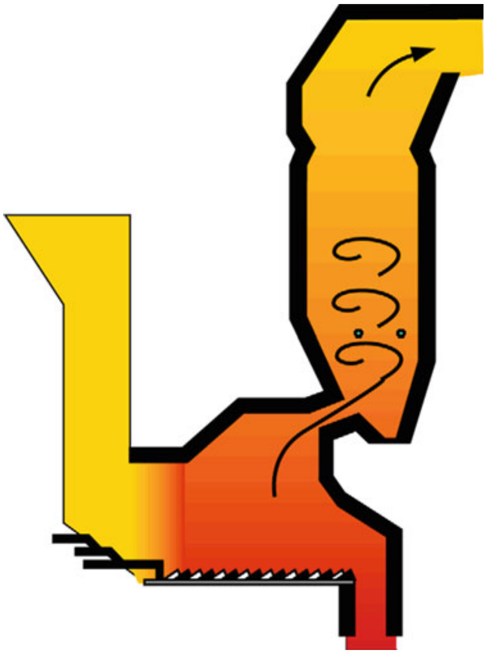
Combustion Control

Stringent statutory requirements stipulate that the combustion process on grate systems be automatically monitored, regulated, and controlled as far as possible. These tasks are fulfilled by the combustion control system.

The purpose of the combustion control system is to ensure that the combustion process takes place under constant conditions whereby:

- Combustion gases and bottom ash burn out completely
- Flue gas emissions are minimized
- A uniform steam flow is generated

A large number of parameters must be taken into account when controlling a combustion system.



Martin Waste-to-Energy Technology. Figure 18 Vortex post-combustion chamber in T3 plants Trieste/Cagliari

Consequently, only experienced operating staff or, if all available process parameters have been taken into consideration, an essentially automatic combustion control system can achieve stable control.

The goal of modern combustion control systems must, therefore, be to manage the inevitable input fluctuations even at the start of the process, i.e., as soon as the fuel reaches the combustion grate. However, because fluctuations in waste quality (water content, amount of combustible matter, heating value of the combustible matter) cannot be noticeably influenced, an attempt is made to estimate these before fuel starts combusting.

Stable and uniform combustion depends largely on the following factors:

- Uniform, continual feeding (fuel bed on the grate)
- Combustion adjusted to waste quality (combustion air flows, distributions, grate speed, etc.)

A stable combustion process (in terms of the position of the main combustion zone, burnout, heat release, etc.) results when the above factors are managed optimally.

MARTIN Infrared Combustion Control (MICC)

The MARTIN Infrared Combustion Control (MICC) is an innovative, modern combustion control system, developed for the reverse-acting grate and the reverse-acting grate Vario. It is a very flexible, extensible, and independent system that is to be integrated in conventional overall plant control systems. The modular architecture facilitates the use of operating mode concepts, thereby enhancing the functionalities associated with classical combustion control. The purpose of this system is to enable optimization of the entire plant in order to meet different operator-specific requirements such as maximum fuel throughput, maximum service period or maximum energy recovery (in the form of heat or electricity), or minimum gas emission (e.g., NO_x).

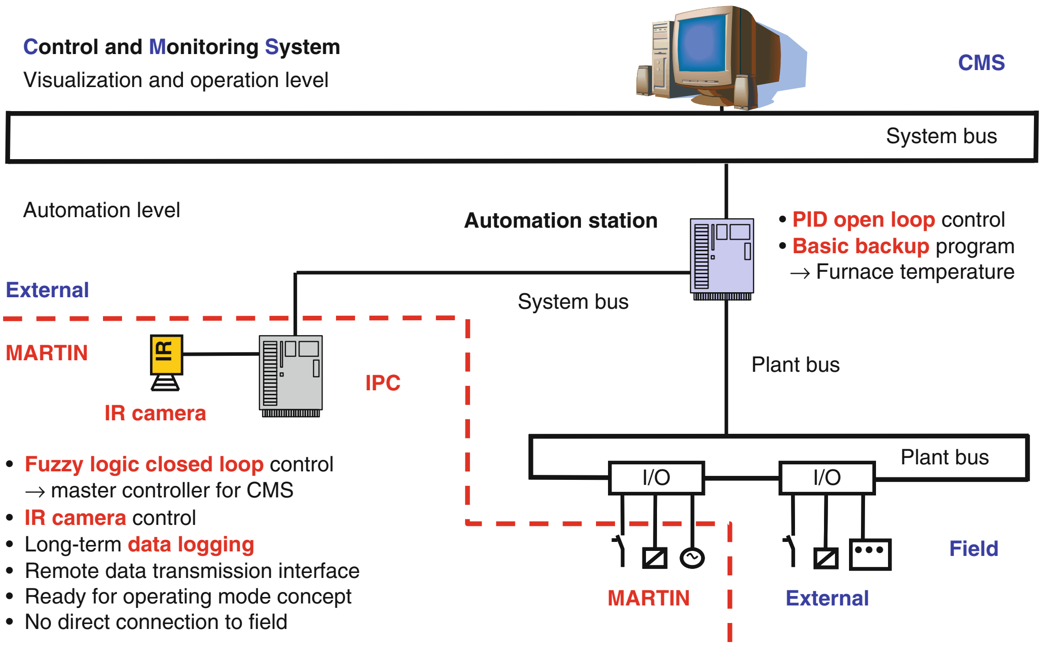
Moreover, additional customer requirements can be incorporated, ideally without having to intervene in the overall plant control system. The MICC system comprises hardware and software. All field devices are connected to the overall plant control system to ensure end-to-end visibility between the operator stations and the field (Fig. 19).

The hardware is based on a globally available high-end industrial PC and is installed in a control cabinet. The software comprises various functional modules:

- Fuzzy logic control of the combustion system
- Infrared camera, image analysis, and signal generation for additional optimizing controllers
- Operating mode concept
- Operational data logging and visualization

Other functional modules such as SNCR control can be integrated in the MICC system in line with market requirements.

The term “combustion control” normally includes both the open-loop and closed-loop control functions for the combustion system and grate. Specific closed-loop control system know-how is implemented in the MICC system. The functions of the open-loop control system and a semiautomatic control system (steam flow control) are programmed in the overall plant control system. Consequently, plant operation can be sustained even while maintenance or optimization work is being performed on the MICC system. With the standardized interface and screens generated in the overall plant control system, the closed-loop system



Martin Waste-to-Energy Technology. Figure 19

MICC integration in plant control and monitoring systems (reverse-acting grate/reverse-acting grate Vario)

can be adjusted from the operator consoles in the main control room without requiring expertise in the MICC component.

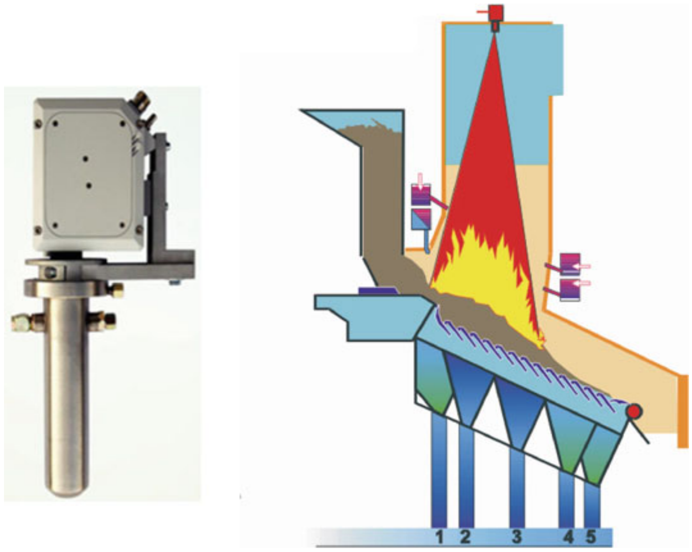
The MICC combustion control system includes the fuel controller that controls feeding of fuel to the grate; the O_2 controller of underfire air flow for the grate; the overfire air controller that controls flow and distribution of overfire air; and the grate speed controller. Additional controllers receiving signals from the IR camera system influence the distribution of underfire air, ram feeder, and grate movement.

There is a choice of control modes: "steam flow," "furnace temperature," or "steam flow/IR temperature." The quality of steam flow/IR temperature control is significantly improved by the IR camera controller. The combustion controllers are implemented as fuzzy controllers. The actuating variables calculated by the combustion control system are transmitted to the overall plant control system, where they are further processed.

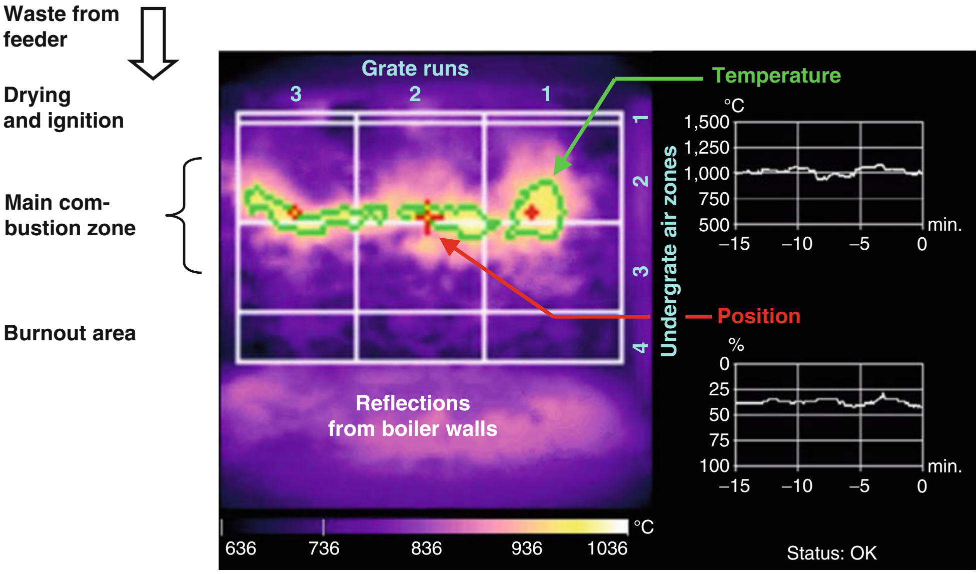
The basic advantage of fuzzy control is its ability to find the "best compromise." Particularly when combusting waste, the process (combustion, boiler, flue gas path, etc.) produces partly contradictory or inexact information for the control system. Fuzzy control processes such information and finds the best solution at that time. Manual intervention is significantly reduced, and on the whole control is very stable. The "if ... then" formulation of control behavior allows every conceivable control case to be formulated simply, which is not possible to the same extent with classical PID control. Compared with classical PID control, the control logic requires less programming, although more complex logical connections can be implemented.

MICC uses specialized infrared camera technology at selected bandwidths. An infrared camera that records the temporal and two-dimensional behavior of the fuel bed surface temperatures from the boiler roof is used to obtain precise additional information from the combustion process (Fig. 20).

Depending on the size of the furnace, the evaluation area of the infrared camera includes (longitudinally for each grate run) grate zones 1–3 up the middle of zone 4, which covers most of the drying, ignition, and combustion stages, and also the entire grate width, which can consist of several grate runs. The infrared camera produces thermographic images of the surface temperature of the fuel bed (Fig. 21).



Martin Waste-to-Energy Technology. Figure 20
Infrared camera (photo/position)



Martin Waste-to-Energy Technology. Figure 21
Image of IR camera

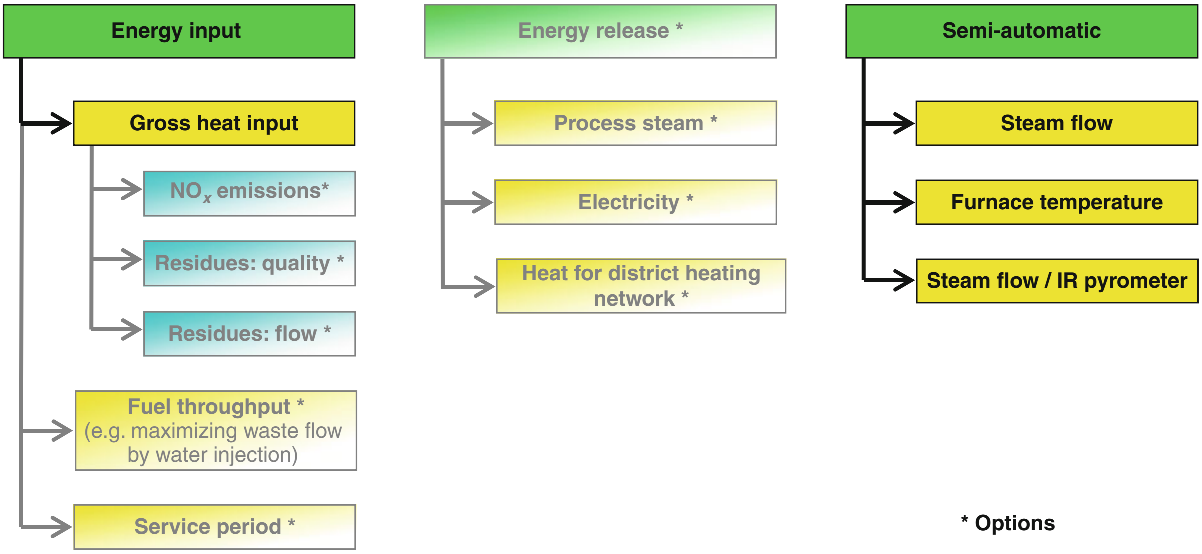
The information delivered by the infrared camera is processed by a specially developed image analysis program using sophisticated mathematical algorithms. Signals for controllers are calculated and transmitted to the combustion control system [14–17]. Operators can see the temporal and two-dimensional distribution of the fuel bed surface temperatures as well as the ash caking distribution on the visible areas of the boiler wall; and also the overfire air nozzles on a separate monitor in the control room. The observer learns to interpret the infrared images and consequently the temperature distribution over time, as well as the overall combustion behavior on the grate. Further optimization measures can then be introduced ahead of the furnace, such as better mixing of the waste in the bunker, etc. The visual information provided to the operators results in more stable and flexible plant operation.

Modern WTE plants are operated flexibly and many are integrated into power plant parks. This means that the most cost-effective mode of operation may vary from case to case. Different goals may be targeted, such as maximum fuel throughput, maximum energy recovery, minimum pollutant emissions, maximum service period, etc. The “operating mode concept” module supports the different modes of operation required by the different goals (Fig. 22).

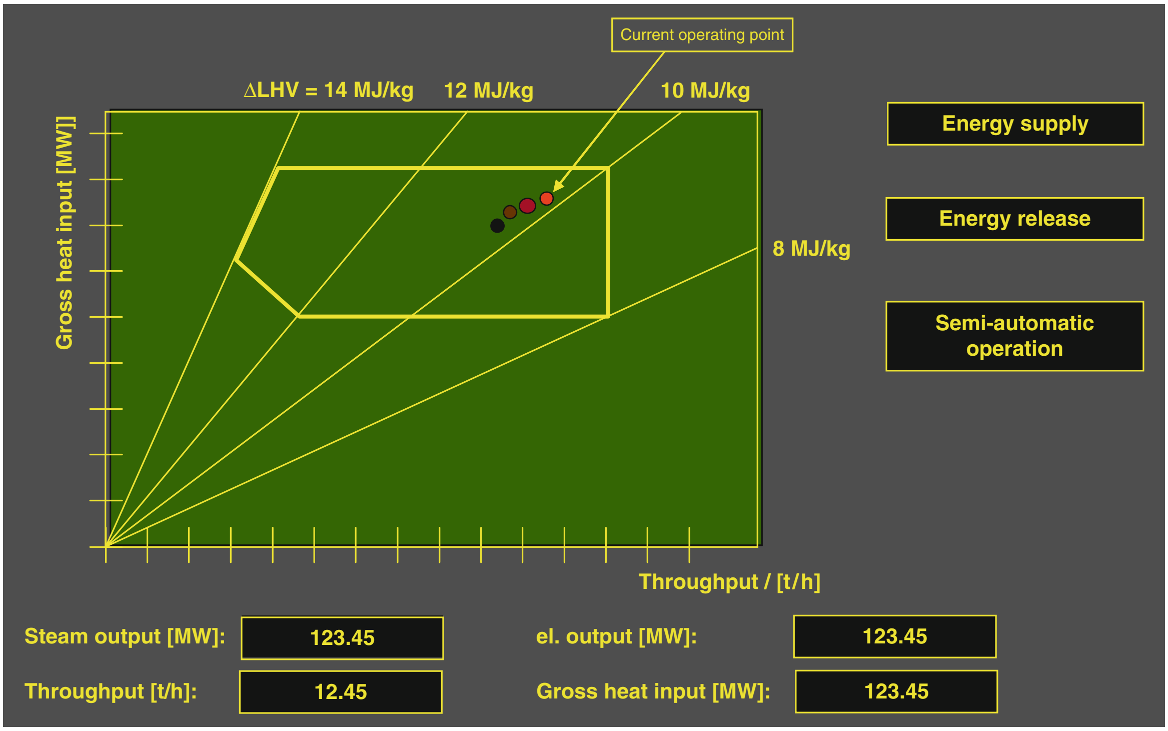
The energy input control concept includes the gross heat input operating mode. Here the heating value of the waste and the waste flow are calculated on the basis of various factors: the goal is to automatically regulate the maximum gross heat input as a function of the waste quality and plant condition. The semiautomatic operating mode is a simplified mode for the combustion system, additionally available in all MARTIN plants. There is a choice of control modes: steam flow, furnace temperature, or steam flow/IR pyrometer temperature control. Semiautomatic mode allows the plant, e.g., to be shut down systematically if important components fail, or allows operation to continue with reduced output. It is possible to switch smoothly between operating modes.

The stoker capacity diagram is displayed in the overall plant control system so that the current operating point is visible (Fig. 23). The operating point is indicated by means of a trailing pointer.

In addition to the available modules for the operating mode concept, further modes may be implemented in response to market demand, e.g., the “energy release” module: It will be possible, in the energy input/gross heat input mode, to have a positive influence on individual parameters such as NO_x emissions, residue quality (bottom ash), or residue quantities. The energy input/fuel throughput mode is designed to maximize waste throughput. The limits of the stoker capacity diagram will not be reached in the energy input/service period operating mode. In this mode, it is more important to achieve maximum plant operating time. Plant shutdown for purposes of overhaul is postponed for as long as possible when this operating mode is selected. The energy release/process steam, energy release/electricity, and energy release/district heating operating modes permit additional selection of the appropriate mode of operation to maximize plant profits. The best



Martin Waste-to-Energy Technology. Figure 22
Diagram of operating modes



Steam output [MW]: 123.45

el. output [MW]: 123.45

Throughput [t/h]: 12.45

Gross heat input [MW]: 123.45

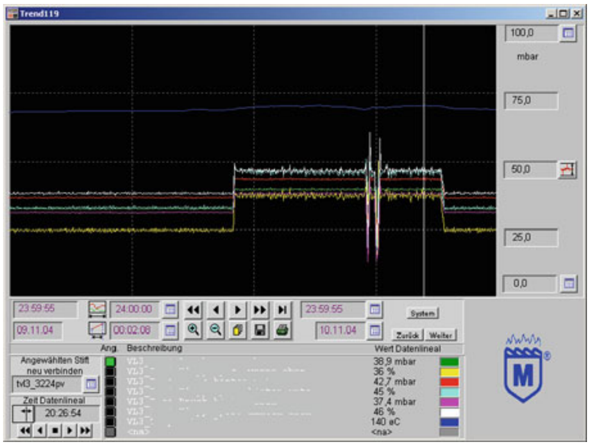
Martin Waste-to-Energy Technology. Figure 23
Stoker capacity diagram

mode can be selected and a constant amount of energy can be supplied to the grid depending on the time of day or year. These operating modes require partial expansion of the fuzzy control system and further observation and start-up phases will be necessary for new parameter settings and optimization measures.

The "operational data logging and visualization" module logs and displays all operating data relevant to the combustion system (Fig. 24). If malfunctions occur or damage is noted, the logged data help identify and eliminate the cause.

The most significant benefits of the MICC component are:

- High flexibility of combustion control
- No dependence on the overall plant control system
- End-to-end visibility from the overall plant control system to the field devices
- Operator strategy can be changed at short notice
- Plant operation in accordance with the plant operator's requirements
- Highly integrated hardware components
- Few interfaces
- Low training costs
- High efficiency
- High level of availability



23:59:55 24:00:00 23:59:55

09.11.04 00:02:08 10.11.04

tvl3_3224pv

20:26:54

Ang.	Beschreibung	Wert Datenlineal	
🟩	VL3	38,9 mbar	🟩
■	VL3	36 %	🟨
■	VL3	42,7 mbar	🟥
■	VL3	45 %	🟦
■	VL3	37,4 mbar	🟪
■	VL3	46 %	⬜
■	VL3	140 øC	🟦
■	<na>	<na>	⬛

Martin Waste-to-Energy Technology. Figure 24
Data acquisition

- Prompt and direct support
- Long-term sustainability through updates

Advanced Combustion Control (ACC)

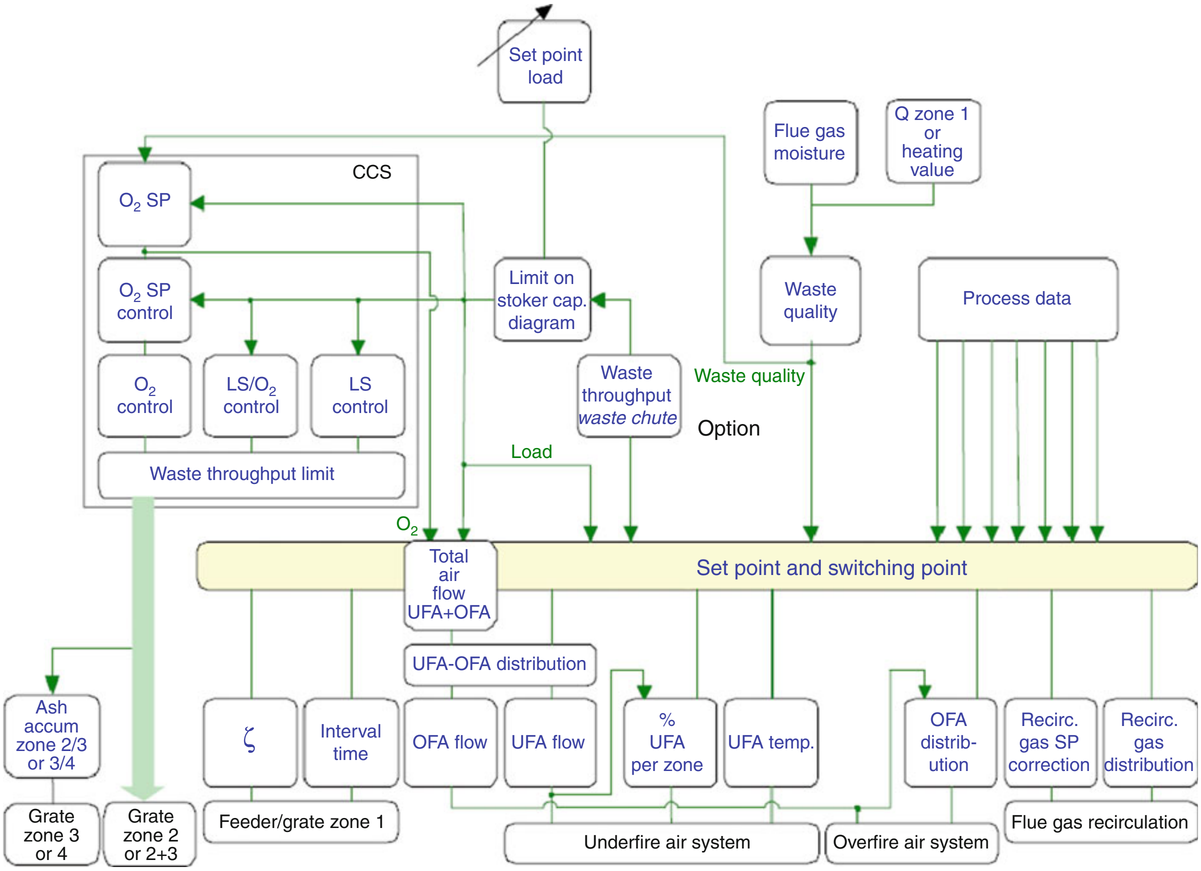
The horizontal grate uses Advanced Combustion Control (ACC) that can be easily retrofitted to existing plants without additional monitoring equipment.

The ACC consists of four control systems:

- Waste input
- Combustion system output (boiler load or O_2 excess)
- Burnout/calcination
- Combustion air

Figure 25 illustrates the basic principles of the ACC system and shows the actuating and controlled variables.

The waste input area (feeder) and grate zone 1 are used for waste supply and waste transport only. They control loading (fuel bed thickness = flow resistance) of grate zone 2 (constant fuel bed thickness) by providing a uniform supply of fuel to the combustion system, and continuously supply an optimum fuel flow to the fire. The long residence time of the waste in the high-temperature area allows it to dry and become combustible. Consequently, even wet waste can be combusted without any noticeable negative effects. To keep the fire in zone 1 to a minimum in the event of very combustible waste, the underfire air is significantly decreased in the first zone. The waste flow introduced via the feeder to zone 1 is controlled using a three-variable (stroke length, idle time, run-on time) control system with a continuously operating feeder.



Martin Waste-to-Energy Technology. Figure 25
Advanced combustion control system (ACC, horizontal grate)

The grate zone 2, or in some cases grate zones 2 and 3, directly controls the combustion control system. Either live steam flow or excess O_2 in the flue gas at boiler outlet is used as the reference variable. The percentage O_2 in the flue gas at boiler outlet or the live steam flow can be controlled with a constant combustion air flow via the waste mass flow using the stoking and transporting motion of grate zone 2, or zones 2 + 3.

Grate zone 3, or grate zone 4, is operated so as to increase the residence time of the bottom ash on the grate and, therefore, the burnout quality. If bottom ash transport is slowed down in zone 3, or 4, ash accumulates between zones 2 and 3, or 3 and 4, and its residence time in the furnace is increased.

The total combustion air flow is specified as a function of load and correctively controlled as a function of two modes:

- Live steam control, for which the O_2 control deviation is the controlled variable
- O_2 control, for which the live steam control deviation is the controlled variable

Load, O_2 in the flue gas, and total combustion air flow always correspond in this control system. The air flow can therefore only be influenced if the O_2 set point is changed.

The total combustion air is distributed between underfire and overfire air accordingly at a specified percent distribution. The specified underfire air percentage is determined automatically as a function of waste quality and load.

The underfire air is distributed to the air zones under the grate as a function of load and waste quality. The air flow is reduced drastically in the first zone if the waste is extremely dry. The air not needed for drying the waste is used as overfire air. In this way, a constant air flow is provided to the combustion process in zones 2 and 3 at a steady load.

In an automatic combustion control system, the set points for all control and regulating systems are calculated as a function of the waste quality and current load and then are taken over by the ACC system. Only minimal intervention is required on the part of the operators; the combustion system adjusts continuously to the current operating conditions. CO and NO_x emissions are reduced by this mode of operation and operational effort is reduced. The quality of bottom ash burnout becomes better and more uniform. Automatic adjustment of the controller settings to suit current “waste quality” and load ensures optimized combustion control. As a result, a very stable output and good flue gas and bottom ash burnout can be achieved with minimal operational effort.

Energy Recovery

Thermal treatment of waste produces energy that is used to generate electricity, process steam, or heat for district heating. The first step is to evaporate water and generate steam. The combustion system and steam boiler must be appropriately matched. As early as 1964, MARTIN directly integrated the steam boiler into the combustion system in the Rotterdam and Paris – Issy les Moulineaux – plants. This concept was further developed and used for WTE plants worldwide.

The arrangement, size, and dimensions of the heating surfaces, i.e., radiation chamber, superheater, evaporator, and economizer must be designed carefully. The large volume of furnace and radiation passes result in low flue gas velocities and relatively long residence times. Furthermore, the type and quality of the ceramic lining in the combustion chamber must be specified. The scope of supply includes recommendations regarding the arrangement and type of online cleaning facilities for the heating surfaces and specifications for the measurement and control devices. High availability and long service periods are consequently achieved.

The boilers in the concepts described below are waste-fired water tube boilers with natural circulation; they are designed to generate superheated steam. Depending on the space requirements and desired cleaning system, there are two typical boiler types to choose from:

- Horizontal boiler
- Vertical boiler

In both types, the first radiation pass is made up of the combustion furnace located directly over the grate and the radiation chamber above the furnace. The furnace begins at the grate surface and ends at a height of about 12 m. The membrane walls of the furnace are designed in the lower area to meet the geometric requirements of the reverse-acting or horizontal grate. Overfire gas nozzles (for overfire air and/or recirculated flue gas) are arranged in the lower area of the front and rear walls for post-combustion and intimate mixing of the volatile flue gas constituents escaping from the grate area.

Ceramic lining is applied to the furnace to protect the boiler tubes against excessive corrosion and to maintain a flue-gas residence time of 2 s at a flue-gas temperature of at least 850°C, as is required in Europe. Studs or anchors suitable to the type of lining used are applied to the wall. The ceramic lining consists of silicon carbide–molded bricks or cast refractory, depending on the location and stress to which the bricks or refractory will be exposed. The walls of the radiation chamber directly over the furnace are protected against corrosion and erosion by applying a weld overlay of nickel alloy (e.g., Inconel 625).

Both boiler types have been designed in many plants for the superheated-steam parameters 40 bar and 400°C. For mixed designs, experience is available for values up to 60 bar and 450°C.

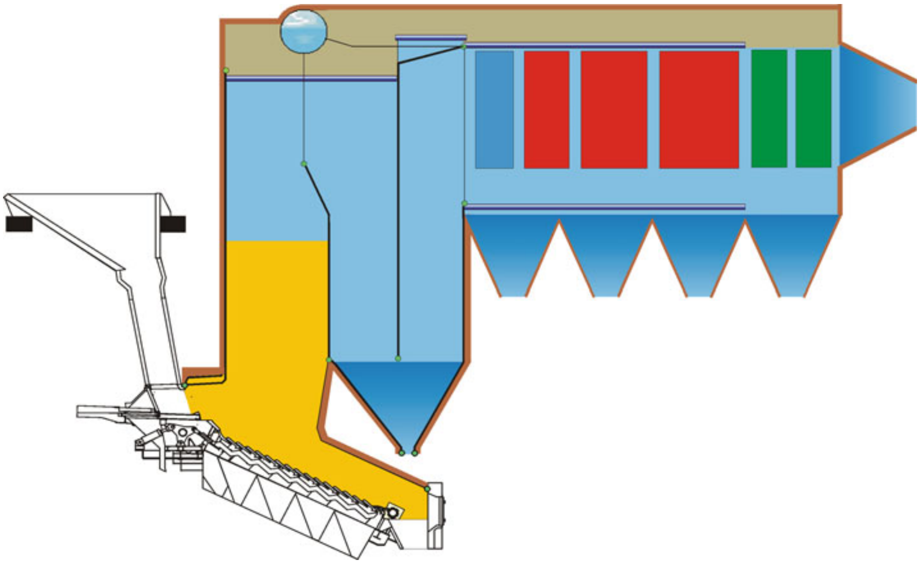
Horizontal Boiler

The horizontal boiler (Fig. 26) consists of three vertical radiation passes and a horizontal pass to accommodate the convective heating surfaces. The three radiation passes are of the same width. The horizontal pass can be up to 30% narrower in order to increase the flue-gas velocity. In the lower area of the second and third radiation passes, the walls are shaped to form a hopper for separating fly ash from the flue gas. The walls of the three vertical radiation passes, the side walls and the roof of the horizontal pass are of the welded, gas-tight tube-fin-tube type (membrane walls). Risers and downcomers connect the membrane walls and convective evaporator(s) in the horizontal pass to the boiler drum, which is arranged transversely across the boiler, forming the natural-circulation evaporator section of the system.

The horizontal pass has convective heating surfaces in the following sequence:

- Evaporator 1 (known as the "cooling trap")
- Superheater (three-stage, with two desuperheaters)
- Evaporator 2 (if required)
- Economizer

All convective heating surfaces consist of aligned bare tubes and the flue gas side is cleaned by means of rapping devices. The energy of the impacts made by the mechanically or pneumatically driven hammers causes the vertically suspended heating surfaces to vibrate,



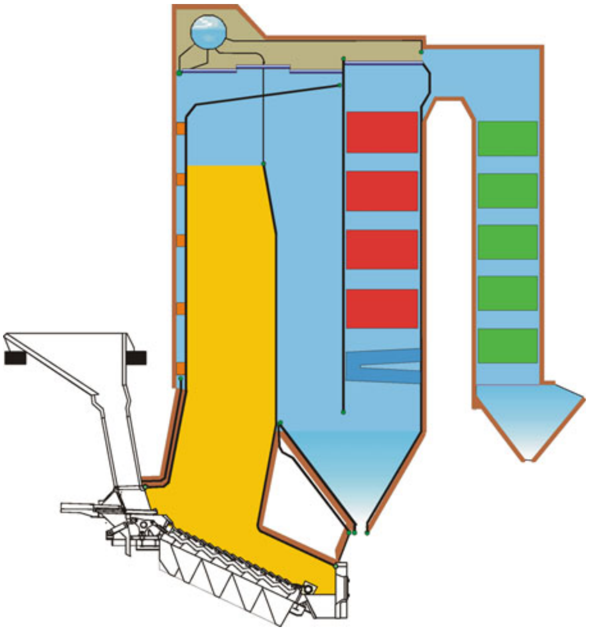
Martin Waste-to-Energy Technology. Figure 26
Horizontal boiler

thereby freeing them of any fouling. The upper heating surface attachment is elastic to prevent cracks due to constant vibration. The bottom sections of the lower headers are specially designed to handle the higher mechanical stress. The economizer is arranged in a gas-tight steel casing and, depending on the space parameters, can either be integrated in the horizontal pass or located in a separate vertical pass. However, a vertical economizer pass requires a different cleaning system (e.g., sootblower or shot cleaning).

Hoppers for collecting fly ash are located below the horizontal pass. The number and form of the hoppers is such that the ash is transported away easily and the flue gases do not unintentionally flow under the heating surfaces.

Vertical Boiler

The convective heating surfaces of the vertical boiler (Fig. 27) are arranged in two vertical boiler passes. The surrounding walls and roofs of the first three vertical passes are of the welded, gas-tight tube-fin-tube type (membrane walls). The convective heating surfaces also consist of aligned bare tubes and are arranged similarly to the convective heating surfaces of the horizontal boiler. However, they are tailored to flue-gas-side cleaning by means of sootblowers due to the tube pitch and depth of the heating surface banks. The economizer is arranged in a welded, gas-tight steel casing and forms the last boiler pass with a hopper for separating fly ash.



Martin Waste-to-Energy Technology. Figure 27 Vertical boiler

The height of the vertical boiler is largely determined by the size of the convective heating surfaces of the third pass. Under the same boundary conditions, this means that the height is greater than that of the horizontal boiler but that the area required is smaller.

Both boiler types are suspended in a supporting steel structure that allows free thermal expansion downward for all pressure-carrying parts. The transition from the boiler to the stationary reverse-acting grate or horizontal grate is designed accordingly.

Energy Efficiency

Because of concerns about climate change and increasing fuel prices, efficient utilization of energy derived from waste has become more significant. Municipal waste has characteristics that make it particularly suitable for the generation of heat and power. Waste is generally available close to the location of heat and power consumption in towns and densely populated areas. MARTIN investigates and evaluates methods and concepts for increasing efficiency by optimizing the combustion system and water-steam circuit using practice-oriented models for preparing large-scale implementation [18].

The standard WTE technology in Europe consists of grate-based combustion systems. Typically, these, sometimes quite old, plants produce 546 kWh of electricity per Mg of waste, which corresponds to a gross energy efficiency of 18% referred to the gross heat input from waste and additional fuels (basis: heating value of 10.44 MJ/kg and electricity production only). Due to in-plant consumption of an average of 150 kWh/Mg of waste, this results in an average exported electricity of 396 kWh (net efficiency of 13%). Most recent WTE plants use steam parameters of 40 bar/400°C. Typically, these plants produce 650 kWh of electricity per Mg of waste, which corresponds to a gross energy efficiency of 22% (heating value of 10.44 MJ/kg). With an in-plant consumption of 150 kWh, this typically results in exported electricity of 500 kWh (net efficiency of

17%). This data refers to the Best Available Technology document on waste incineration by the EU IPPC directive (BREF) [19].

In some European countries, landfilling of municipal waste is restricted and efforts are being concentrated on further improving the energy efficiency of WTE plants beyond the values mentioned above. The driving force behind the implementation of high-energy systems is usually a premium for renewable electricity from waste.

There is a large potential for improving the use of the energy contained in municipal waste. On the one hand, waste can be diverted from landfilling, on the other hand, the energy efficiency of WTE plants can be improved. In this respect, the main topics apply to power generation: steam parameters (pressure and temperature of superheated steam), flue gas heat losses (temperature at boiler outlet, excess air rate), steam condensation conditions (air or water condensers), thermal cycles (intermediate superheating, external superheating, two or three pressure systems), and in-plant consumption (SNCR/SCR, excess air rate).

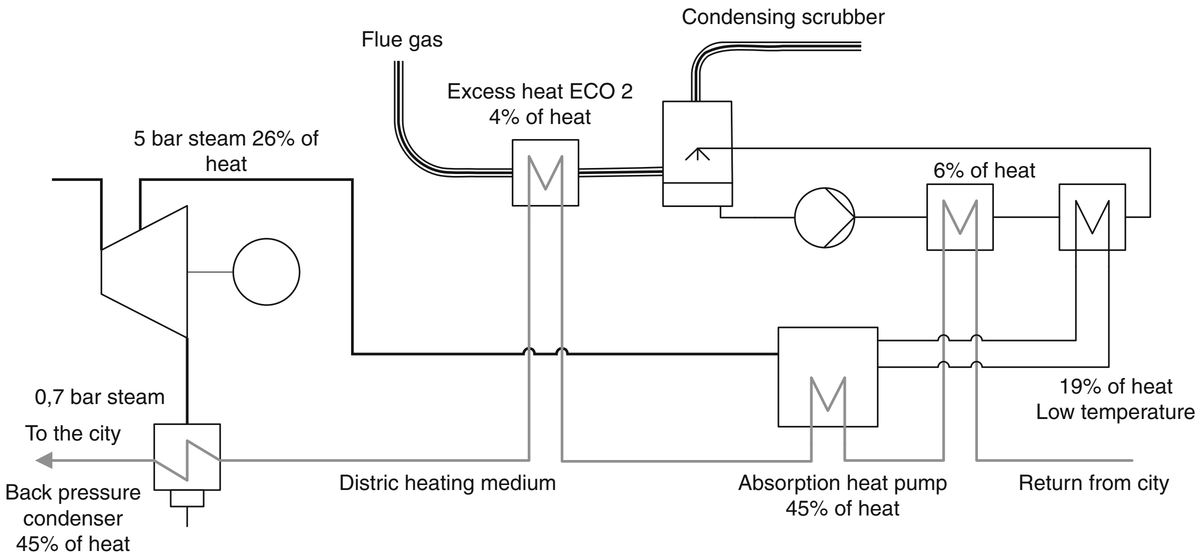
Examples of recent innovative WTE plants with MARTIN grate technology and highly efficient power generation can be found in Brescia (IT), Amsterdam (NL), and Bilbao (PT). The Brescia plant has an increased gross efficiency of produced electricity of 27% through increased steam parameters, reduced flue gas losses and minimized in-plant consumption. The new plant in Amsterdam achieves 30% with additional intermediate superheating and water condensers. A further increase in energy efficiency is then only possible by external superheating with natural gas in combined cycle plants, as in Bilbao. However, innovations also took place in the field of heat recovery. The Malmö plant (SE) is an example, where efficiency has been increased by using heat generated from flue gas condensation for district heating. Twence (NL) is another example, where a high degree of energy recovery is achieved by combining heat and power production.

In the case of power generation, the main limitation on increasing energy efficiency is posed by the increased cost and corrosion risk. For the use of heat, climatic limitations and the cost of district heating grids are important considerations. In Europe, potential exists to increase the proportion of WTE to over 10% of the overall renewable energy produced as half of the energy contained in municipal waste is of biogenic origin [20].

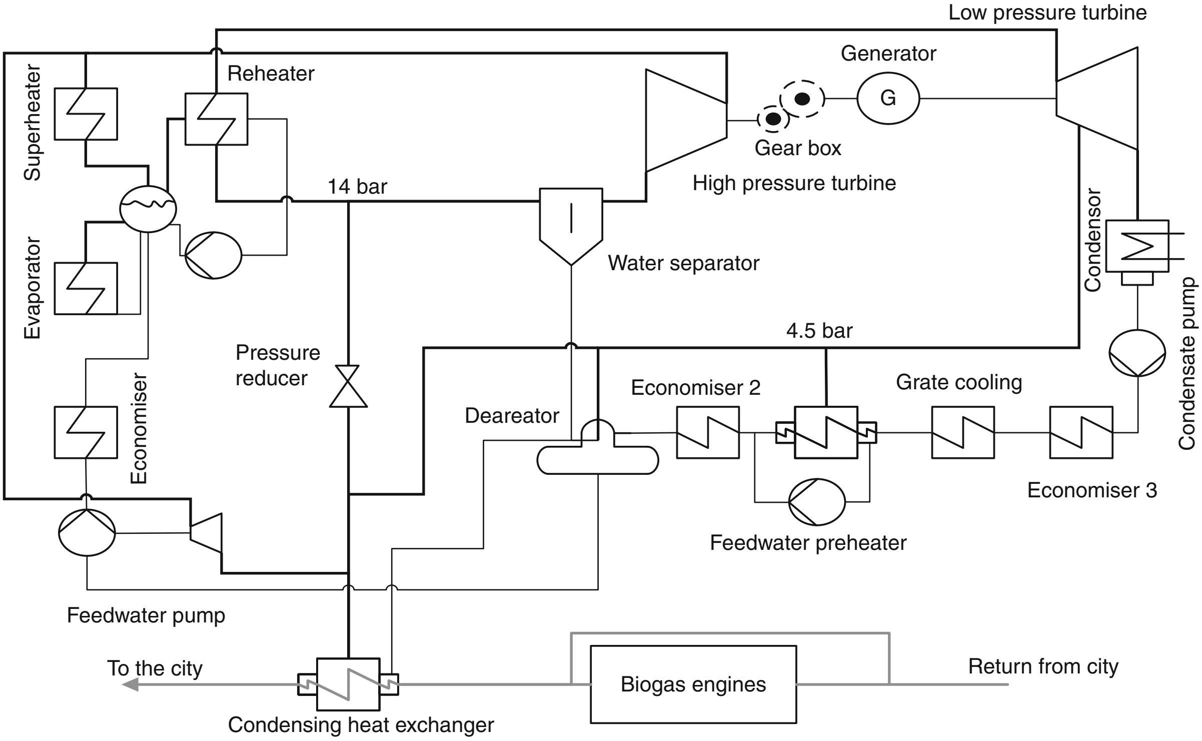
Energy performance indicators are plant-specific figures which allow a power/throughput-independent comparison and classification of power plants with respect to their energy efficiency. Efficiency is generally defined as the ratio of useful output and input. In WTE plants, the input is waste and the useful output can be electricity, heat, or even recovered materials. Gross electric efficiency, net electric efficiency, and thermal efficiency are most commonly used. It is essential to define clear system boundaries when comparing the efficiency values of different plants.

In Gothenburg (SE), where the focus is on the production of heat, a large district heating system is installed, where the WTE plant is only one of many heat suppliers. During the summer, the demand for heat is further increased by absorption chillers distributed around the city. The district heating system's feed and return temperature depends to a large extent on the outside temperature. The feed temperature varies from 75°C to 110°C, whereas the return temperature lies between 40°C and 55°C. The yearly averages are 80°C to 45°C. Figure 28 shows the four stages of the heat production system. Due to the low return temperature, it is possible to use the low-temperature heat of the condensing scrubber directly. In the next step, more heat from the scrubber water is extracted by means of absorption heat pumps. These heat pumps are driven by steam extracted at the 5 bar tapping point of the turbine. Depending on the operating conditions, excess heat from the second economizer is used in the following stage. The last stage comprises the back pressure condenser of the turbine operating with floating pressure, depending on the required feed temperature, to maximize electricity production. Nearly 30% of the produced heat is extracted directly from the flue gas, which reduces the loss of electricity production and thereby decreases the power loss coefficient [21].

The new WTE plant in Amsterdam (NL) is designed for maximum electricity production. With a sophisticated water–steam cycle (Fig. 29), the plant achieves an electric efficiency of more than 30%. The turbine can only be operated with elevated pressure of 130 bar at low live steam temperature with a water separator and external reheater system. In 2008,



Martin Waste-to-Energy Technology. Figure 28
Heat production Gothenburg (SE) [21]



Martin Waste-to-Energy Technology. Figure 29
Plant diagram of new WTE plant, Amsterdam (NL) [21]

the plant was retrofitted with a condensing heat exchanger to produce heat too. The district heating medium is heated up from the return temperature of 50°C to 120°C (design feed temperature) in two stages. In the first stage, waste heat from biogas engines installed in the plant is used to heat part of the district heating medium to 105°C. Upstream of the condensing heat exchanger, the hot stream is mixed with the bypass stream resulting in a temperature of 61°C. Steam extracted from the turbine is condensed in the condensing heat exchanger to reach the final temperature of 120°C. Most of the steam is extracted from the 4.5 bar tapping point, but since the plant was not designed for heat production, not enough steam can be extracted for full load operation of the heat exchanger. Steam from the 14 bar tapping point has to be reduced to 4.5 bar to be used in the heat exchanger. The biogas engines are able to provide 3.5 MW_{th}, while the heat exchanger of the WTE plant is designed for a maximum of 18.3 MW_{th} heat output [21].

Process Simulation

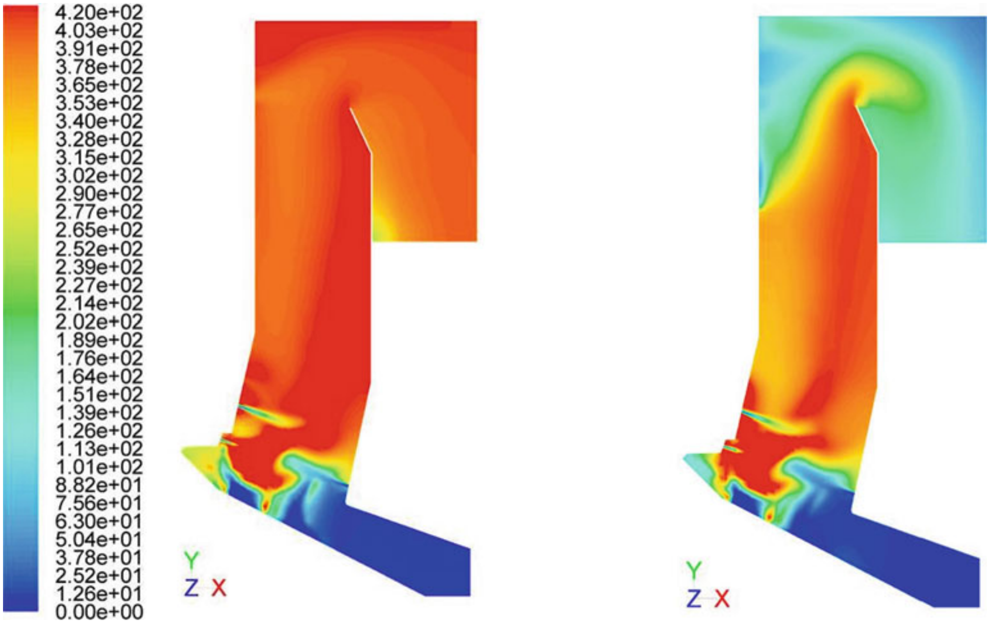
A wide range of simulation tools are used in the WTE sector for the design of new plants as well as for the investigation of operating issues. These tools apply thermodynamic representations of the water–steam cycle to more complex representations such as CFD modeling.

A Boiler Dynamic program is used as a basis for the combustion and boiler/water–steam cycle simulation of a power plant. This program code is designed for engineering complex heat exchangers, e.g., steam boilers and power stations. It allows the user to build a graphical schematic representation of a plant in a graphical user interface by selecting predefined power plant elements from a library and to simulate their operation. Mass and energy balances are calculated for each element, which is considered to be like a black box. The program code makes it possible to simulate different types of boilers such as: natural circulation boiler, force flow boiler, Benson boiler (supercritical), WTE plants, fire tube boilers, fluidized bed combustion etc. It includes static and dynamic modules. The first module will be used to plan a new power plant, to study an existing plant regarding fouling, to analyze the influence of different modifications (fuel, operating point, changing of elements such as heat exchangers). The dynamic modules are useful to analyze the behavior of the boiler in special cases such as start-up, load changes, or shutdown [22].

More complex problems in boilers such as detailed combustion reactions, three-dimensional temperature distributions or formation of emissions are analyzed by means of Computational Fluid Dynamics (CFD). Numeric modeling and simulation to describe the flow processes in waste combustion are regarded as state-of-the-art and are in successful, widespread use. For combustion problems, it has been shown that this software gives accurate results and takes all aspects such as chemical reactions, heat transfer, fluid flow, and thermal radiation into account. CFD works by solving the equations of mass conservation, momentum conservation, and energy conservation over a region of interest, with specified conditions on the boundary of that region.

For boiler design purposes, it is enough to provide a general model including combustion, heat transfer, and radiation. For combustion reactions, there is a wide range of chemical reaction models starting from simple ones for fast chemistry such as the eddy dissipation model or more complex ones for finite rate chemistry such as the eddy dissipation concept model, considering chemical kinetics. For the purpose of investigating unburned carbon monoxide in the flue gas, models with finite rate chemistry are needed, whereas for heat release and heat transfer calculation, simpler models are sufficient to give realistic results.

CFD is also used to investigate more complex problems such as the formation of NO_x (Fig. 30). The CFD software includes numerical models of NO_x reactions from literature for the three NO_x formation mechanisms (fuel, prompt and thermal NO_x). These models can be used to calculate the formation of NO_x for every type of mechanism, showing the different areas in the boiler where this emission is formed. For reduction of NO_x on the other hand, the SNCR mechanism is implemented and can be selected. Ammonia and urea can be injected in gaseous, liquid, or solid form into the boiler to react with NO_x emissions. Different injection models with varying geometries can be chosen to simulate introduction of the reducing agent as realistically as possible [23]. However, these models are used to



Martin Waste-to-Energy Technology. Figure 30
CFD modeling: NO_x/NO_x – SNCR calculation (NO_x [mg/Nm^3, referred 11% O_2, dry])

optimize the SNCR system and consequently reduce NO_x emissions and additive consumption.

These injection models are further used to investigate the trajectories of fly ash particles and can help to predict deposits or unfavorable flow patterns. Combining particle tracking with other codes, it is possible to numerically observe the growth of the deposit and thereby caused changes in flue gas flow.

A basic precondition for CFD simulation of the gas phase of a boiler is the provision of exact initial values by means of a suitable fuel bed model [24]. A wide range of models is available. This is due to the fact that many different effects occur on the grate bed, which can be modeled in different ways and are considered important by some, while others neglect the same effect. Ensuring transparency and understanding the fuel bed model used are significant when working with CFD. Therefore, it is often easier to develop a new model which is designed to comply with one's needs instead of using other models with a limited capacity for adaptation and tailored for some other purpose.

Instead of modeling the total fuel bed model with motion, gasification, and pyrolysis of the waste, it is sometimes easier to just define the heat or species release along the waste bed to calculate starting values for CFD. These models are more empirical and give no answer to what is happening on the grate, but have proven to be good enough for boiler design calculations or even research topics in the gas-phase combustion [25].

Of primary interest in simulation are the practical implementation and therefore the validity of the model calculations. Validation is not of great importance for thermodynamic water–steam cycle calculation since these models are based on energy and mass balances and therefore give good results. CFD simulations on the other hand are based on a wide range of models simplifying reality. Choosing the right model is important to get realistic results. If there is any possibility, CFD simulations are always validated by measurements. In WTE plants for validation, a mass and energy balance over the boiler is combined with temperature measurements in the flue gas path to perform accurate validations of CFD calculations. For this purpose, grid measurements of temperature are performed in the range of the SNCR level to obtain a two-dimensional temperature profile, which then is compared with the temperature profiles calculated by the CFD software.

In a project covering overall validation on an industrial-scale WTE plant, the development of a fuel bed model yielded improvements in the basic boundary conditions for CFD simulation. Comprehensive measurements of temperatures, flue gas concentrations, and velocities were performed and compared with simulated values. Using the model approaches selected, it was established that the model for CO burnout and for a number of other simplified partial models, such as heat dissipation in the convective area of the heat exchangers, did not permit accurate analysis or forecasting for these subareas. However, good matches were achieved, in particular with regard to temperature and flue gas concentration values for CO_2 and O_2. There were some mismatches in the speed components due, above all, to the turbulence prevailing in the first pass of WTE plants. Nevertheless, it was possible to identify tendencies correctly [26, 27].

NO_x Reduction

Low NO_x Technologies

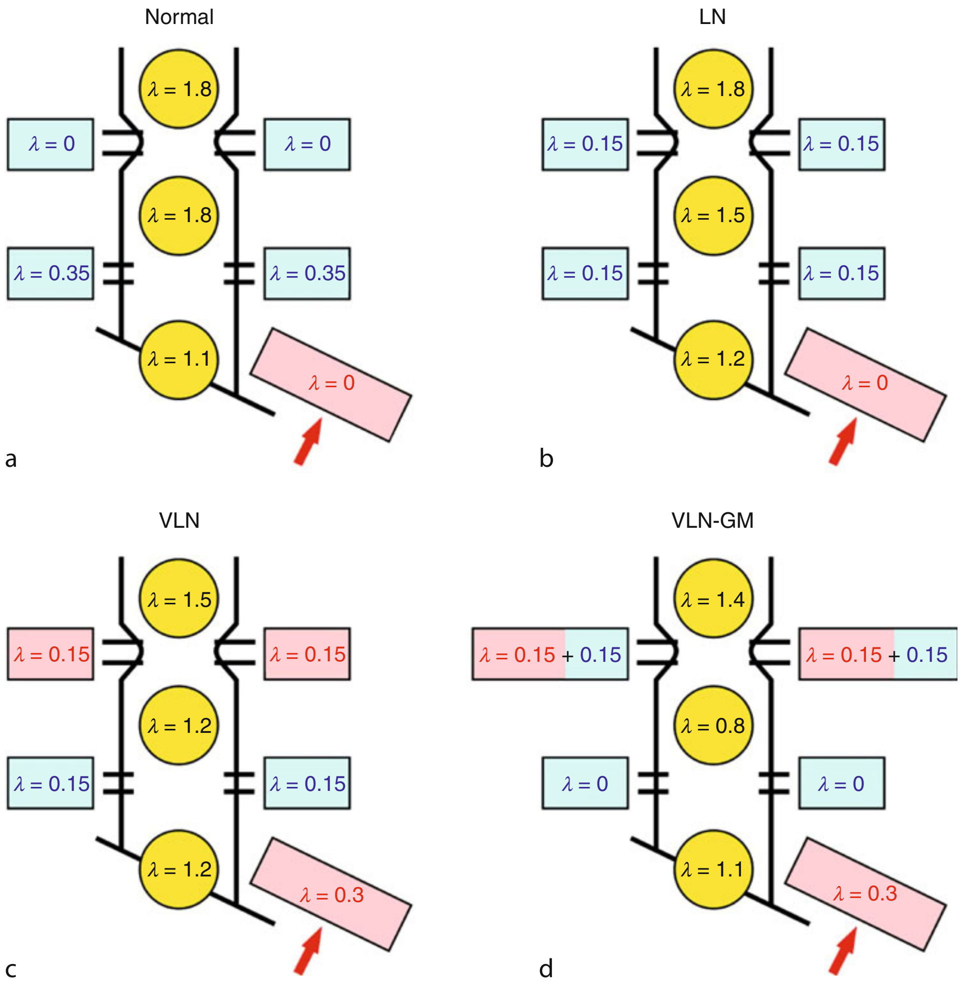
Most of the waste's nitrogen content is transferred to the flue gas during combustion as nitrogen oxide NO_x. The limit values for NO_x emissions continue to decrease as a result of statutory or regulatory requirements. At the same time, the operators of thermal waste treatment plants are increasingly being put under pressure to reduce investment and operating costs. In the EU, there is a combustion directive that defines a maximum emission limit value of 200 mg/m^3 NO_x as a daily average value referred to 11% O_2. Compliance with this limit value is possible with the SNCR process, which injects ammonia or urea into the furnace. In some cases, SCR catalytic converters are used. However, these involve higher costs (investment/operation) and energy consumptions. Based on the National Emission Ceilings (NEC) defined by the Gothenburg Protocol, it can be expected that the limit values for NO_x in the EU will become even more stringent.

In this respect, various concepts of low NO_x technologies were developed to significantly reduce the NO_x values downstream of the combustion system via primary measures (Fig. 31). This is due to the fact that chemical reactions that convert the primarily formed NO_x back to nitrogen are promoted as a result of the reduced excess air and consequently higher temperatures in the lower area of the furnace.

Figure 31a shows the excess air for a conventional combustion system setting. The underfire air is set to be slightly superstoichiometric. An excess air rate of approximately 1.8 is achieved by supplying overfire air for flue gas burnout. The NO_x content is typically 400 mg/Nm^3. The option in Fig. 31b shows the point at which some of the overfire air is introduced into the upper furnace area being moved. The NO_x content can be reduced to approx. 300 mg/Nm^3. This process option, known as *LN* (*Low NO$_x$*), is extremely suitable for being retrofitted to existing plants. However, it can also be implemented in the design of new plants.

Figure 31c shows the process known as *VLN* (*Very Low NO$_x$*). This process achieves a reduction in the levels of excess air and consequently higher temperatures in the lower area of the furnace by means of internal flue gas recirculation. The internal flue gas recirculation system comprises the extraction of flue gas in the rear area of the combustion chamber and its subsequent use as mixing gas in the upper area of the furnace. This ensures optimal mixing of the flue gases. It has been proven under continuous commercial conditions that NO_x values below 250 mg/Nm^3 are achieved with internal flue gas recirculation. These values are reduced to less than 80 mg/Nm^3 (all NO_x values referred to 11% O_2, dry) by injecting ammonia or urea. A further feature of this process is that a low NH_3 slip is adhered to at the same time. The VLN system is shown and explained in detail in Fig. 32. Both LN and VLN processes have been tested on an industrial scale over longer periods and are offered for new plants, and also within the framework of plant retrofits.

Figure 31d shows a further option, the process known as *VLN-GM* (*Very Low NO$_x$-Gasification Mode*). This option was specially developed for customers preferring gasification to combustion. There is no injection of overfire air in the lower furnace area in this model. Significantly, substoichiometric conditions (gasification) are created in the furnace by simultaneously extracting flue gas in the rear combustion chamber area. Mixing gas and overfire air are supplied to the upper furnace area to ensure complete oxidation of the flue gases.

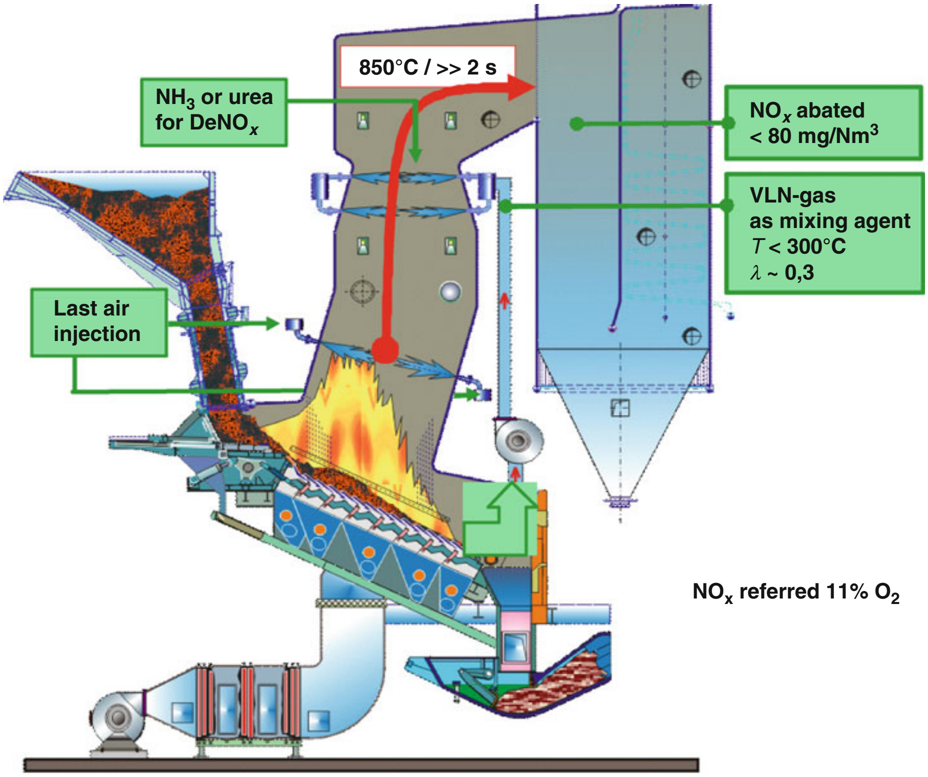


Martin Waste-to-Energy Technology. Figure 31
Various concepts of Low NO_x technologies

The *VLN process* (Fig. 32) is based on a classical grate-based combustion system for municipal waste, where the "VLN gas" is drawn off at the rear end of the grate and is reintroduced into the upper furnace just below the ammonia injection positions. The positive effect of this patented process is twofold: On the one hand, drawing off the VLN gas leads to combustion conditions which promote the inherent NO_x reduction processes such that fuel NO_x is largely reduced to nitrogen. On the other hand, reinjection of the VLN gas cools the flue gases down and enforces their mixing with injected ammonia or urea. This leads to improved efficiency of the SNCR system.

NO_x values of 80 mg/m^3 with an NH_3 slip of less than 10 mg/m^3 have been reached in tests at the WTE plant in Bristol (US). Further industrial-scale tests in Thiverval (FR) and Oita (JP) have confirmed these results. At the Thiverval plant, which has a municipal solid waste throughput of 12 Mg/h, NO_x could be reduced from 190 to 80 mg/m^3 during test operation with the VLN components.

The VLN gas has a temperature below 300°C and is reinjected at a position at which the furnace temperature is around 1,000°C. This typically corresponds to a level of 8–12 m above the grate depending on the capacity of the unit and the type of waste. The overfire air pressures are reduced to around 10 mbar, which is considerably less than in conventional WTE plant design. Nevertheless, superstoichiometric conditions are reached at the overfire air level, which is an


850°C / >> 2 s
NH_3 or urea for $DeNO_x$
NO_x abated < 80 mg/Nm^3
VLN-gas as mixing agent $T < 300$°C $\lambda \sim 0{,}3$
Last air injection
NO_x referred 11% O_2

Martin Waste-to-Energy Technology. Figure 32
Very Low NO_x (VLN) process (NO_x [mg/Nm^3, referred 11% O_2, dry])

advantage compared with air-staged or fuel-staged combustion systems. The residence time from the last combustion air injection at the overfire air level to the 850°C level in the furnace is significantly increased. A further advantage of the VLN system is the reduced flue gas velocity in the lower furnace due to internal recirculation via the VLN duct. This leads to a reduction in the fly ash carried over to the boiler.

The VLN gas is reinjected into the front and rear sides of the narrow section of the upper furnace. This leads to an intensive barrier of turbulence, which reduces the flue gas temperature and blocks the passage of flames or unreacted gases. The test plant results give rise to the expectation that corrosion is significantly reduced in the furnace above the VLN level as well as in the superheaters. On the other hand, temperatures between the overfire air and the VLN level are higher than in conventional combustion and higher grades of furnace protection material should be used there. Another advantage of the VLN system is the reduced excess air rate, which allows cost reduction in the boiler and flue gas cleaning and improved boiler efficiency.

Selective Non-catalytic Reduction (SNCR)

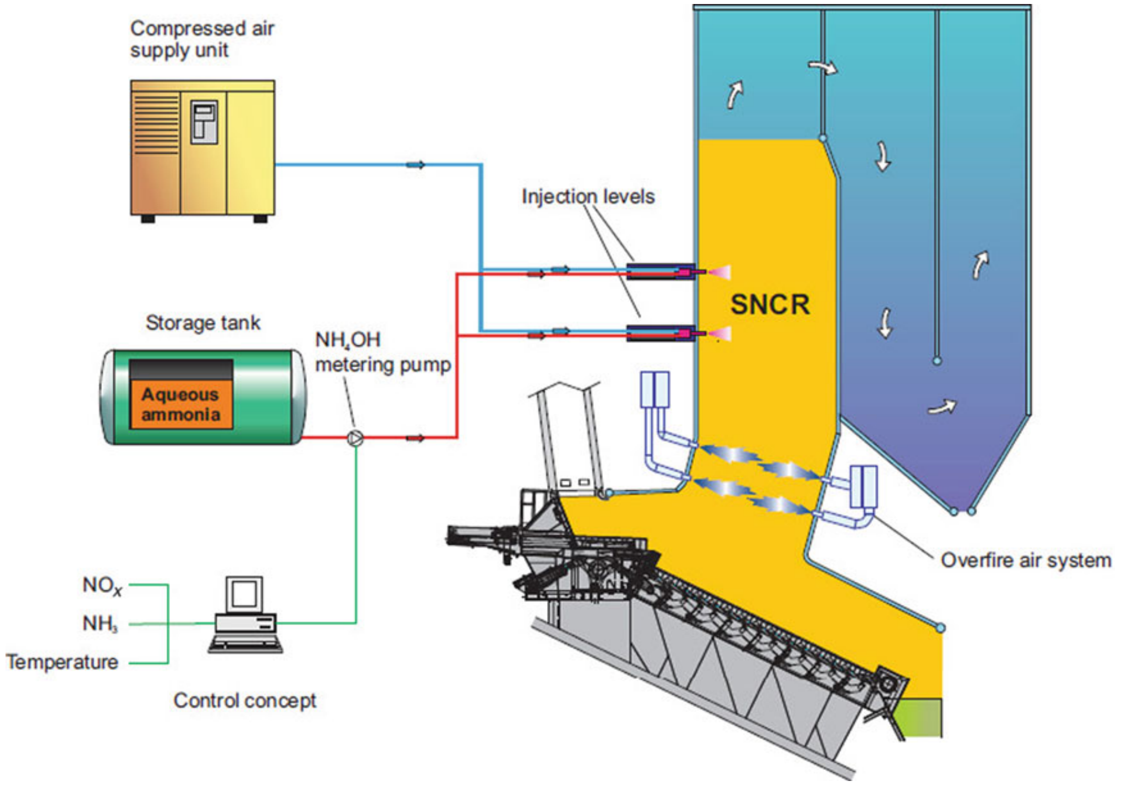
Nitrogen oxides are present in the flue gas due to the high nitrogen content of the waste. Their levels can be reduced, on the one hand, by means of primary measures as the various concepts of Low NO_x technologies and on the other hand by secondary measures as Selective Catalytic Reduction (*SCR*) or Selective Non-Catalytic Reduction (*SNCR*).

The nitrogen oxides are reduced to nitrogen (N_2) and water (H_2O) in the SNCR system, based on the principle of selective non-catalytic reduction (SNCR), by injecting aqueous ammonia (NH_4OH) into the furnace. The aqueous ammonia is injected via nozzle lances at two levels within a temperature range of 850–1,050°C. Moreover, the ammonia can be converted to N_2 or even to additional NO_x in secondary reactions at temperatures above 1,050°C without NO_x participation. Below temperatures of approximately 850°C, the $DeNO_x$ reactions occur only very slowly. Compressed air (on request also softened fresh water or steam) is used as the atomizing medium in order to achieve the finest possible reducing agent distribution.

The compressed-air SNCR system (Fig. 33) requires only one carrier medium (compressed air) for injection and cooling and is adjusted to the specific conditions of the MARTIN combustion system and their flow profiles. Excess ammonia must be supplied due to the dependence on factors such as the temperature window at the injection point, mixing with flue gas and the secondary reactions taking place. In order to remain within the temperature window, the compressed-air SNCR system distributes the aqueous ammonia to two levels as a function of the temperature. This means that there are always two levels in operation at the same time.

The compressed-air SNCR system consists of several system components. The aqueous ammonia storage tank serves as an area for delivery, storage, and drawing off of aqueous ammonia. As specified by regulations, the depressurized storage tank is mostly designed as a double-walled construction. The compressed air supply unit, a screw-type compressor including refrigerant dryer, delivers compressed air continuously. The pumping station includes NH_4OH metering pumps for the upper and lower injection level. Distribution of the mass flows to the individual nozzle rows in the boiler walls takes place in the distributor stations. All lances are required to introduce the aqueous ammonia into the furnace. The compressed air flow ensures an extremely fine distribution of aqueous ammonia droplets.

There are three control concepts that can be optimized using an in situ NH_3 measurement device for



Martin Waste-to-Energy Technology. Figure 33
Compressed-air SNCR system

reducing slip. They can be used to regulate the amounts injected as required and to optimize use of the aqueous ammonia with changing NO_x raw gas concentrations and combustion conditions. In addition, a balance control system steplessly shifts the point at which NH_4OH is injected to the optimum temperature window between the upper and lower injection levels.

NO_x values of up to 70 mg/Nm^3 (referred to 11% O_2, dry) are reliably achieved with a low level of NH_3 slip at the boiler outlet. Features of the SNCR system are:

- Use of cost-effective standard components/easy to replace spare parts
- Stabile spray pattern
- Defined setting of the NH_4OH flow
- Rapid adjustment of the NH_4OH flow when temperature and NH_3 slip change

Within the temperature window, the compressed-air SNCR system can achieve low NO_x emission values. However, the NH_3 slip increases due to the excess aqueous ammonia required for the SNCR reduction method. The system can therefore be equipped with a raw gas catalytic converter. This has the following advantages:

- Use of NH_3 slip for further reducing NO_x
- Only half of the catalytic converter volume is used compared with a complete SCR system in the raw gas area to achieve the same NO_x emission value
- Peripheral plant devices for reheating the raw gas are not required

The NH_3 slip and NO_x emissions remaining in the flue gas flow after the SNCR reaction pass through the raw gas catalytic converter, e.g., downstream of evaporator and upstream of the inlet into the economizer bundles. The converter is equipped with a sootblower system.

At the WTE plant Brescia (IT), a high-dust selective catalytic reduction system has been successfully implemented since 2006 (SNCR + high-dust catalytic converter). The catalytic converter is installed along the gas path where the temperature is already suitable for operation and reheating is not needed. In this way a lower concentration of NO_x (<70 mg/Nm^3) is achieved in the flue gas as well as lower ammonia consumption and lower concentration of ammonia slip (2–6 mg/Nm^3) in the flue gas [28].

Residue Qualities

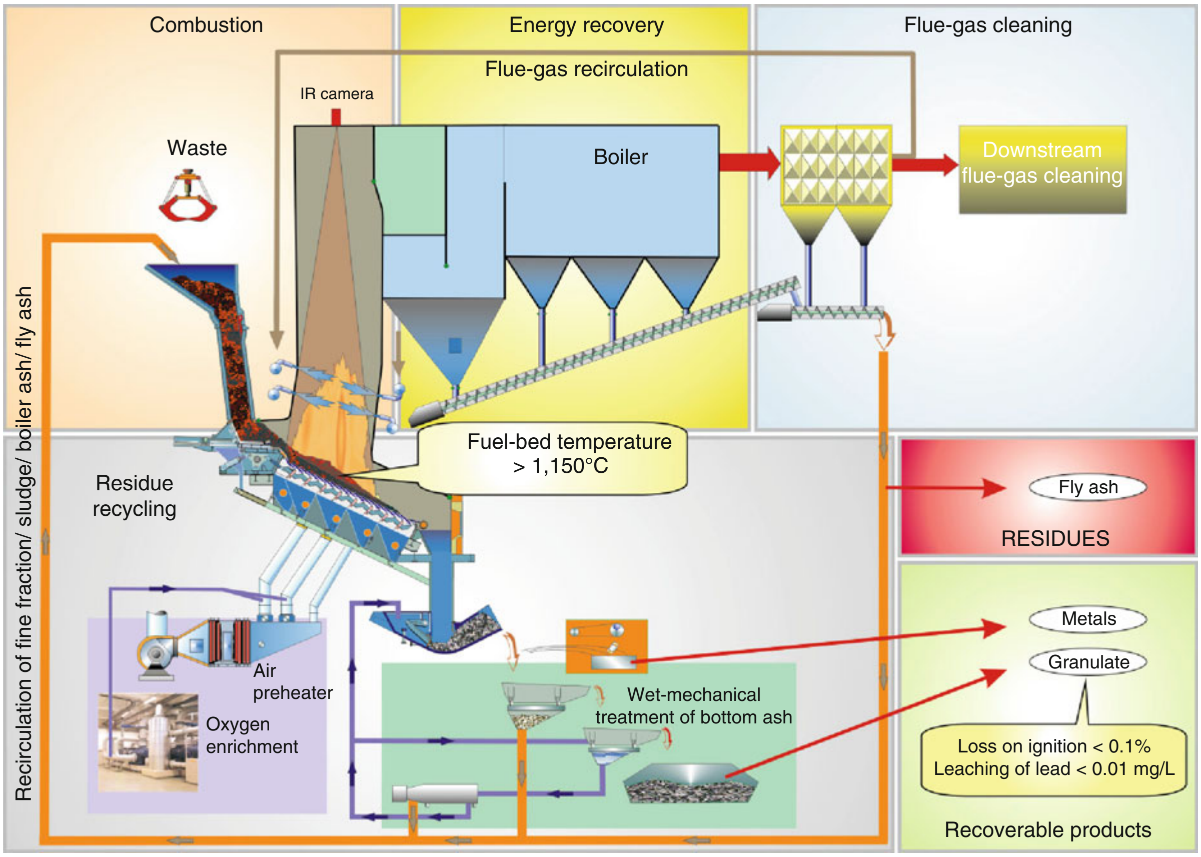
SYNCOM-Plus

Bottom ash produced during thermal waste treatment accounts for the largest mass flow of the waste input at approximately 25% by weight. In Europe, this bottom ash is currently used as a building material, e.g., in road and landfill construction, or as a mining filler, e.g., in salt or coal mines. However, a considerable percentage is sent to landfills.

In the SYNCOM process, as described earlier, combustion air is enriched with oxygen, so that fuel bed temperatures are considerably higher, thereby causing increased sintering of the bottom ash. To further improve bottom ash quality, the SYNCOM-Plus process was developed, whereby a downstream wet-mechanical treatment process to separate a granulate (Fig. 34) is added to the SYNCOM process [29, 30]. The separated fine fraction and sludge as well as the boiler ash/fly ash are then recirculated to the combustion system for further sintering and the destruction of organic compounds at high fuel bed temperatures of approximately 1,150°C (Fig. 35).

A compact large-scale pilot plant consisting of selected units for the continuous wet-mechanical treatment of bottom ash as well as recirculation of fine

Martin Waste-to-Energy Technology. Figure 34
SYNCOM-Plus granulate



Martin Waste-to-Energy Technology. Figure 35
SYNCOM-Plus concept

particles and boiler ash into the combustion system was implemented in the SYNCOM WTE plant in Arnoldstein (AT). The continuously accumulating bottom ash flow coming from a wet-type discharger was first dry screened to separate a fine fraction of <5 mm, then washed and wet screened to separate a granulate of >2 mm and a suspension in a double deck screening machine. Because the wash water was circulated, any particles <2 mm contained therein had to be completely separated in order to minimize the addition of fresh water. A decanter centrifuge was used for this purpose. The sludge that accumulated was fed into the combustion system; the particle-free wash water was conveyed into a storage tank for washing and wet screening. The pilot plant operated continuously and was always directly connected to the bottom ash discharge.

An average of 428 kg/h fine fraction, 180 kg/h sludge, and 58 kg/h boiler ash were recirculated. In total, this amounted to approximately 6.5% of the entire hourly waste input. No significant accumulation of the fine fraction in the bottom ash total as a result of recirculation could be detected after determining the screening curves (comparison of particle size distribution with/without recirculation). Analysis of the combustion parameters (steam, fuel bed temperatures, etc.) and the raw gas measurements at the boiler outlet indicated no significant influence being exerted by SYNCOM-Plus operation. The granulate meets the criteria for solids and leachates laid down by the recovery regulations and all requirements pertaining to aggregates for unbound materials and anti-frost layers for use in civil engineering and road construction.

Figure 36 shows the entire input and output mass flows in the SYNCOM-Plus process for the SYNCOM WTE plant in Arnoldstein (AT). The greatest residue mass flow is the bottom ash output (3.74 Mg/h). Of this, the SYNCOM-Plus process produces 2.82 Mg/h of a granulate with good recovery qualities. Fine fraction, sludge, and boiler ash are recirculated; the filter ash contains additives from flue gas cleaning and must therefore be disposed of.

Figure 37 is a graphic illustration of the energy flow for the SYNCOM WTE plant in Arnoldstein (AT) showing complete conversion of the waste input into electricity and specifying a value of 2.25 MW for in-plant consumption. However, the components of the SYNCOM-Plus process make only a small contribution in this respect. As an option, heat for district heating and process steam can also be generated. However, this has no influence on the in-plant electricity consumption.

The SYNCOM-Plus trials at the SYNCOM WTE plant in Arnoldstein (AT) demonstrated successful results in continuous operation mode and therefore proved that both, wet-mechanical treatment and recirculation, in continuous operation are feasible. No influence on combustion or the raw gas could be detected [31].

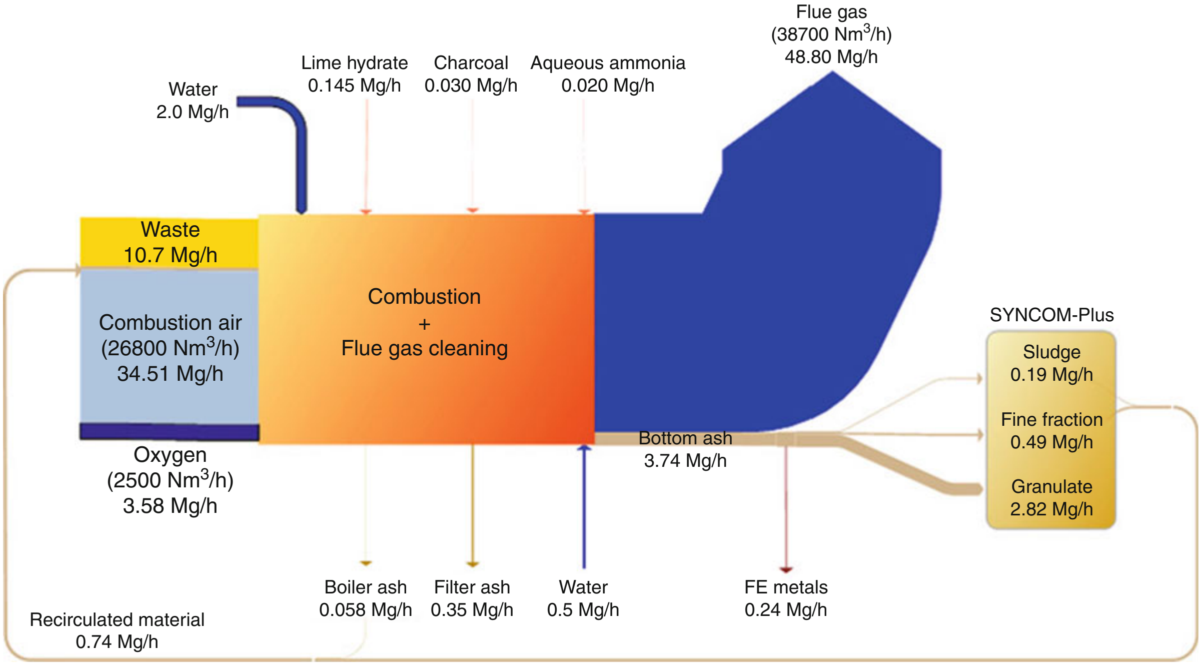
The removed material flows (fine fraction and sludge) and the wash water can be recirculated within the process so that SYNCOM-Plus produces no residues for disposal and the entire process is effluent-free.

The characteristic features of the SYNCOM-Plus process are listed below:

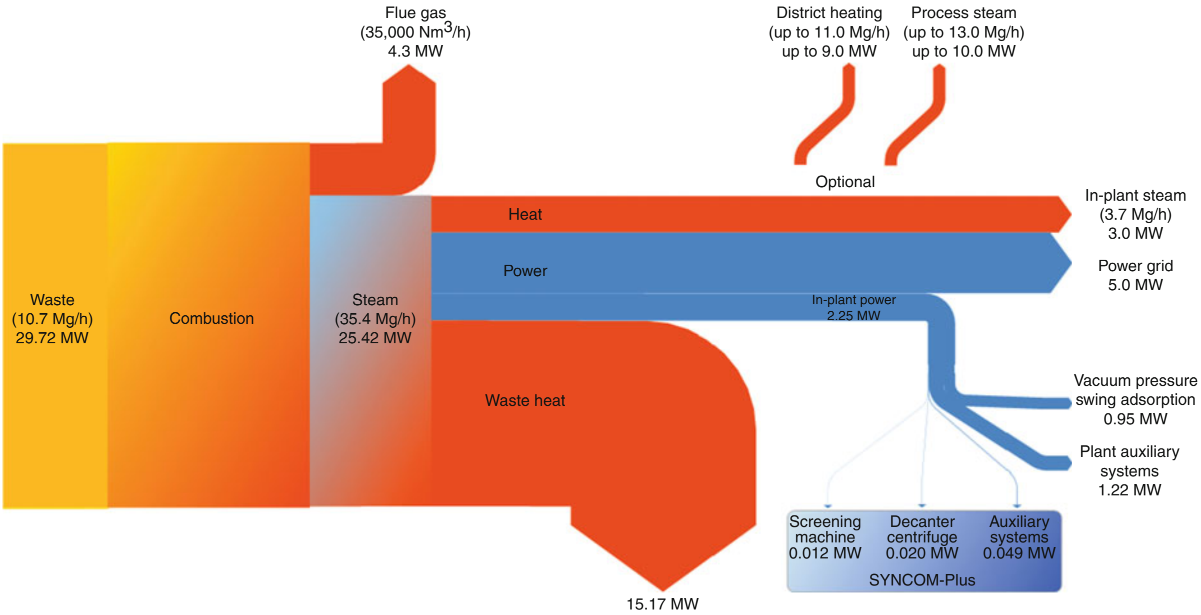
- Wet-mechanical treatment/recirculation feasible in continuous large-scale operation
- Optimized granulate quality with optional recovery
- No vitrification of residues necessary/destruction of dioxins >90%
- Reduced amount of residues for disposal/effluent-free process

Dry Bottom Ash Discharge and Separation

The WTE bottom ash that is discharged from the grate and in particular its metals, can be recycled. Dry discharge of bottom ash offers significant advantages in this respect. Not only are the metals of a better quality, but



Martin Waste-to-Energy Technology. Figure 36
SYNCOM-Plus – mass flow



Martin Waste-to-Energy Technology. Figure 37
SYNCOM-Plus – energy flow

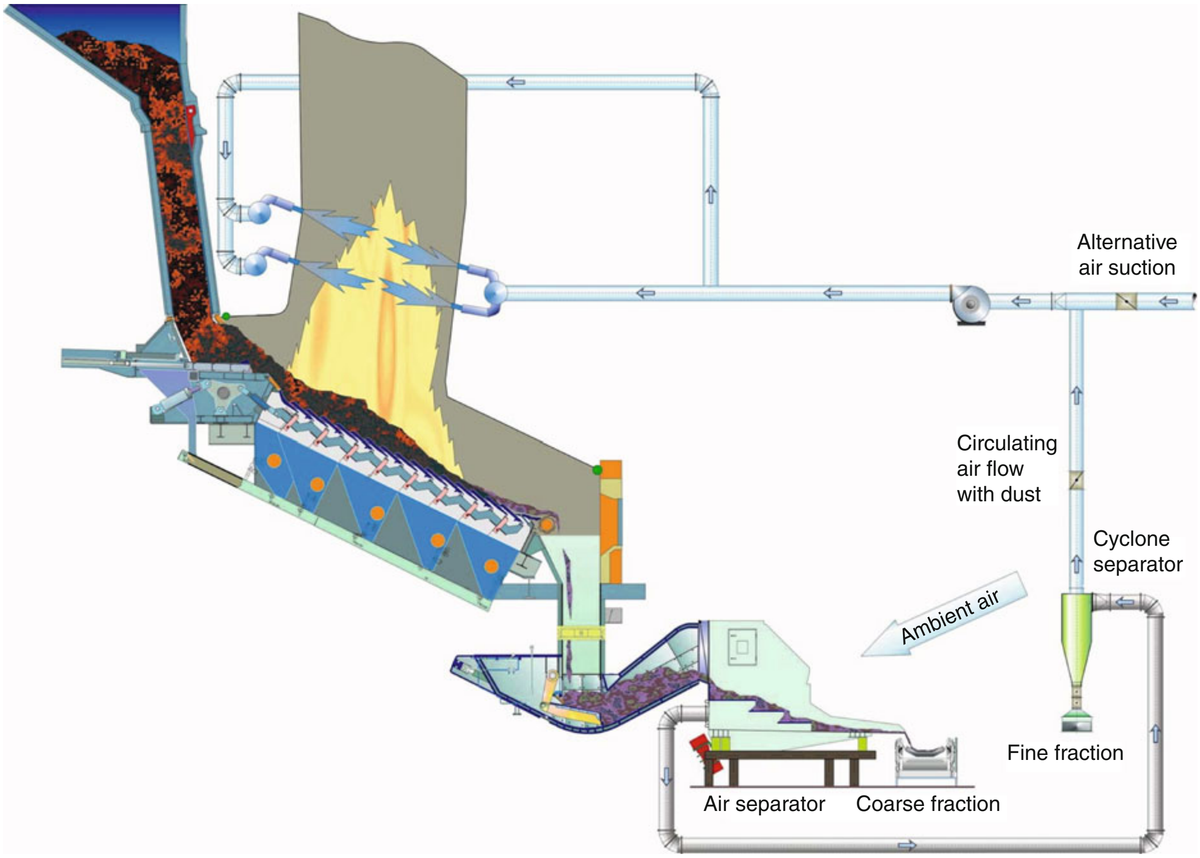
the fine fraction that generally contains heavy metals can be separated from the coarse fraction more effectively.

Dry discharge of bottom ash from waste combustion is becoming increasingly important in the context of recycling raw materials from combustion residues. Particularly in Switzerland, but also in numerous other countries, there has been much interest over the last few years in this process. Discharge of graded, dry bottom ash is on the one hand economically interesting because metal is separated more efficiently, returns from metal recovery are maximized and disposal costs are reduced through reductions in weight and lower transport costs. On the other hand, there are advantages due to the better quality of the dry discharged bottom ash and the fact that subsequent treatment, processing, and recycling is easier.

The simple and robust technology used in the dry discharge system is based on generally known process-engineering separation and transport procedures and uses the advantages of a fractioned bottom ash discharge. The ram-type bottom ash discharger is therefore used in unchanged form but with a newly developed and patented air separator and a cyclone separator (Fig. 38).

For dry bottom ash discharge, the ram-type discharger is operated without water. The bottom ash is discharged in dry form from the combustion system. The dry discharged bottom ash is conveyed directly to an air separator (Fig. 39). The fine fraction and bottom ash dust is extracted in a defined manner. Depending on the extraction speed, the fine fraction and dust particle size is set to values <4 mm. The air separation area is enclosed by a housing, in which negative pressure is constantly maintained, thereby preventing false air from entering the furnace or dust from getting into the boiler house. Similar to wet-type discharge, the surface temperature of the discharger lies in the range 40–60°C in the dry discharge mode.

Essentially, three product streams are separated out of the dry bottom ash: coarse fraction, fine fraction, and bottom ash dust. The latter two are discharged out of the air separator with the air flow and conveyed to a cyclone separator, which ensures that they are separated from the air flow. The unburdened air, containing a minimal amount of residual bottom ash dust, is conveyed in a defined manner to the combustion air system via the overfire air. The fine fraction separated in the cyclone separator is sent to a recycling process or landfilled.



Martin Waste-to-Energy Technology. Figure 38
Dry bottom ash discharge and separation

On the basis of the successful semi-industrial preliminary tests, the operator of the Monthey WTE plant (CH) decided to implement dry discharge systems in both combustion lines for industrial-scale operation.

The characteristic features of the dry bottom ash discharge and separation are listed below:

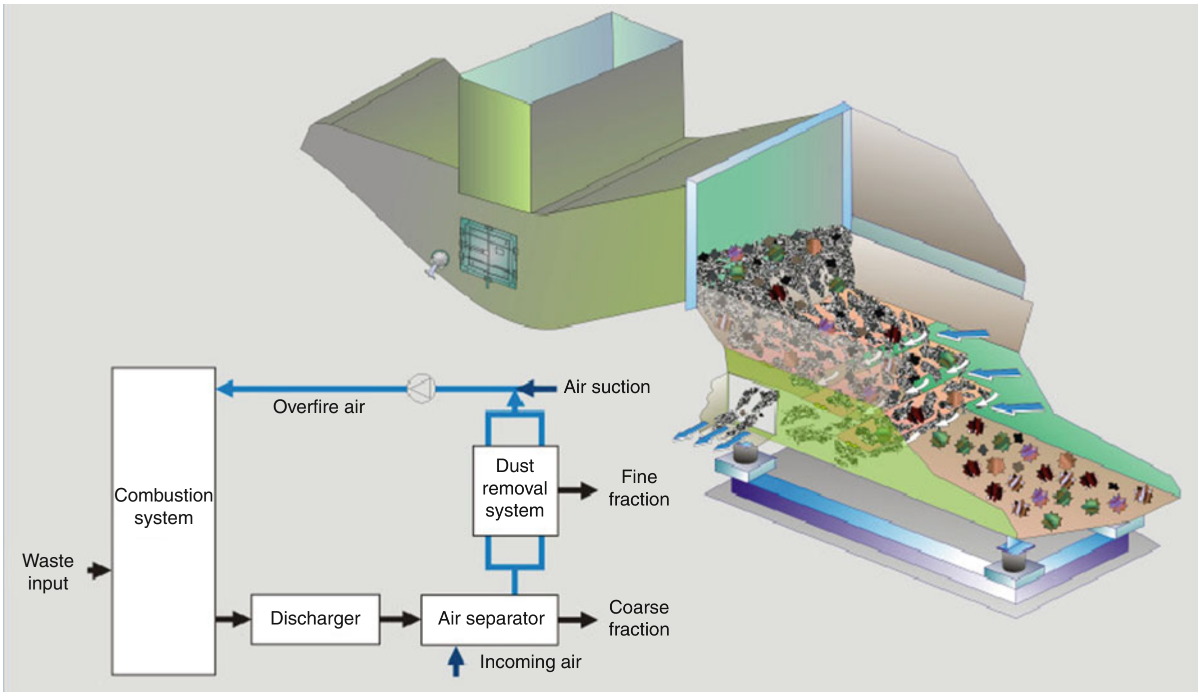
- Weight reduction (no water!)
- More effective metal separation
- Better quality metals
- Lower water consumption
- Better bottom ash quality

Fly Ash Treatment/Recirculation

In view of the continual depletion of raw materials, sustainable processes for the recovery of recyclables are becoming more and more important. The process of selective zinc recovery from the acid-scrubbed fly ash of thermally treated waste is one example of a cost-effective, process-integrated method for recovering economically profitable heavy metals.

By means of this process, cadmium, lead, and copper are separated and the valuable metal zinc, which is contained in high concentrations in the fly ash, is recovered in pure form (Zn $>$ 99.99%). After acidic fly ash scrubbing, the filter ash cake has an extremely low heavy-metal content. Any organic matter that remains in the cake subsequent to scrubbing is returned to the combustion system so that it can be destroyed (Fig. 40) [32].

The synergies associated with the residues occurring with wet flue gas cleaning are used during the process. During acidic ash extraction, the heavy metals in the fly ash are mobilized and extracted by the acidity of the quench water. At the same time, the excess acid content of the quench water is neutralized by the



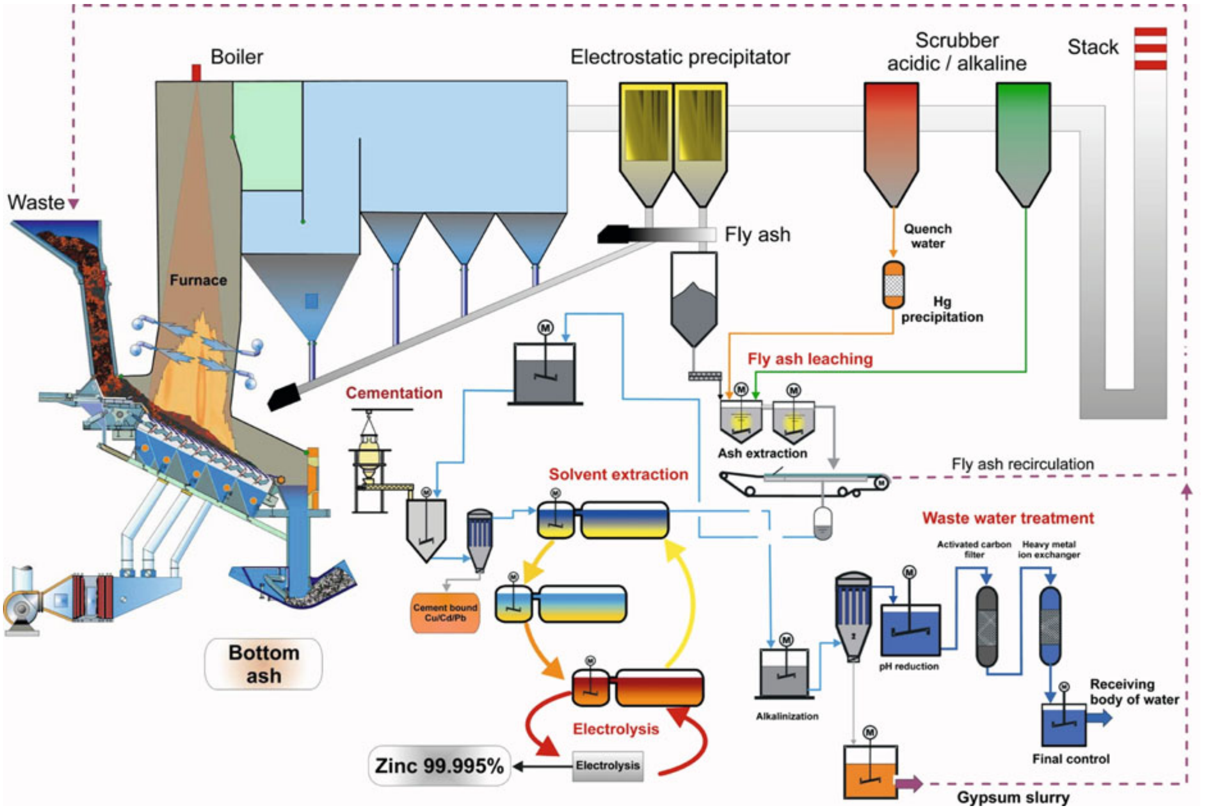
Martin Waste-to-Energy Technology. Figure 39
Concept of dry bottom ash discharge/air separator

alkalinity of the fly ash. In a filtration stage, the filter ash cake, which has an extremely low heavy metal content, is separated from the filtrate containing heavy metals. The filtrate serves as the base material for subsequent targeted heavy metal separation and recovery. Cadmium, lead, and copper are separated using a reductive process and recovered as a metal mixture in lead works. Zinc, which is present in the filtrate in economically interesting concentrations, is separated from the pre-cleaned filtrate using a selective liquid–liquid extraction method, and then concentrated and recovered electrolytically as pure zinc (Zn $> 99.99\%$) [34]. In a waste water treatment plant using lime milk, the filtrate with its extremely low heavy metal content is neutralized, filtered, and freed of all remaining traces of heavy metals using selective heavy-metal ion exchangers. The resulting clean waste water can be supplied directly to the receiving body of water. After filtration, a small amount of residual metal sludge consisting largely of gypsum and alkaline earth metal hydroxides remains and is disposed of.

To destroy the organic matter in the filter ash cake, in particular the dioxins and furans, the filter ash cake with its extremely low heavy-metal content is returned to the combustion system. As proved in several industrial-scale trials at WTE plants, recirculation of the filter ash cake has no effect on plant operation, clean gas parameters and the quality of the resulting bottom ash.

By combining the individual process steps, the residues from flue gas cleaning can be sustainably and efficiently treated. Using the processes described here, thermal waste treatment results in bottom ash and metal fractions containing recyclables that can be selectively returned to the raw materials cycle.

In addition to the sustainable recovery of secondary raw materials and closing of substance cycles, this process also has economic advantages. Value is added through the recovery of zinc and significant cost savings are made in relation to the treatment of waste water with low heavy-metal contents, thereby ensuring the cost efficiency of the overall process. The process could consequently achieve sustainable, ecologically relevant, and economically interesting recovery of residues from flue gas cleaning systems in thermal waste treatment plants.



Martin Waste-to-Energy Technology. Figure 40
Fly ash treatment/recirculation/Zn recovery [32, 33]

Together with its cooperation partner, MARTIN has implemented research and development into treatment processes and the recirculation of treated fly ash in industrial-scale WTE plants.

Reference Plants

In Figs. 41–46 some examples of WTE plants using MARTIN technology are provided. A current reference list is available in [1]. In order to provide solutions for customers with small waste volumes, modular and standardized small-scale plants have been specifically developed. These plants are suitable for waste throughputs of 2.5–8 Mg/h per combustion line.

Future Directions

Sustainability, recycling, resource conservation, and global warming are the greatest global challenges facing us today and will continue to occupy our attention in the years to come. Waste combustion is an essential component of all modern waste management concepts and plays a central role in the handling of these issues, which will become increasingly important as time goes on.

Experience and statistics from many countries clearly show that the combustion of residual waste does not conflict with the avoidance, recovery, and recycling of waste in any way. The rate of recycling of secondary raw materials is particularly high in countries predominantly using combustion, while their rate of landfilling of residual waste is low.

There is no doubt that CO_2 is produced and released into the atmosphere with the flue gases generated during waste combustion. However, when assessing the relevance of these emissions to the climate, a difference must be made in terms of where the CO_2 originates. Approximately 50% of this CO_2 is biogenic in origin (paper, cardboard, wood, materials, leather, etc.) and

Reverse-acting grate
SYNCOM

Number of lines:	1
Waste capacity per line:	10.7 Mg/h
Thermal capacity per line:	29.6 MW
Steam output per line:	35.2 Mg/h
Steam pressure:	40 bar
Steam temperature:	400°C

Martin Waste-to-Energy Technology. Figure 41
Arnoldstein (AT)

Reverse-acting grate

Number of lines:	2
Waste capacity per line:	15.0 Mg/h
Thermal capacity per line:	38.3 MW
Steam output per line:	46.0 Mg/h
Steam pressure:	35 bar
Steam temperature:	242°C

Martin Waste-to-Energy Technology. Figure 42
Vienna Spittelau (AT)

Reverse-acting grate	
MSW	
Number of lines:	2
Waste capacity per line:	23.0 Mg/h
Thermal capacity per line:	88.3 MW
Steam output per line:	115.0 Mg/h
Steam pressure:	60 bar
Steam temperature:	450°C
Biomass	
Number of lines:	1
Waste capacity per line:	23.0 Mg/h
Thermal capacity per line:	100.0 MW
Steam output per line:	115.0 Mg/h
Steam pressure:	73 bar
Steam temperature:	480°C

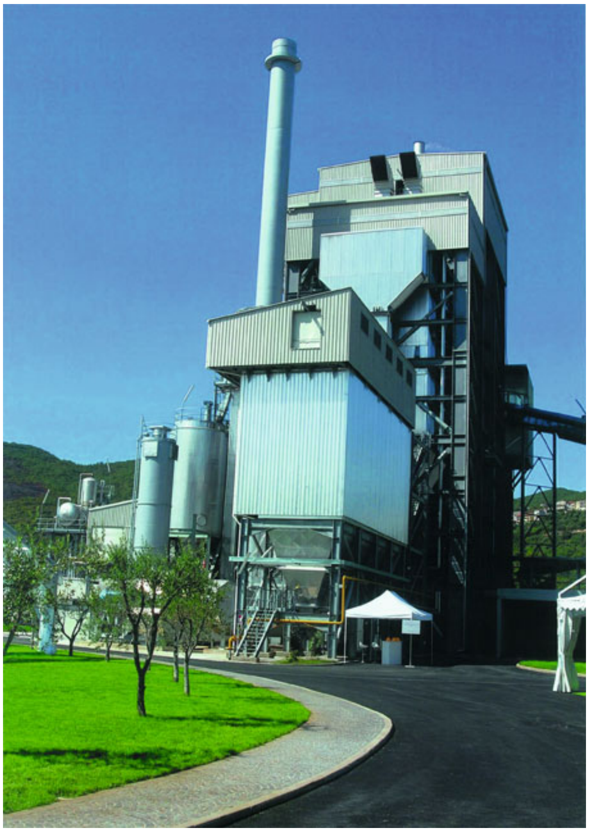
Martin Waste-to-Energy Technology. Figure 43
Brescia (IT)

Reverse-acting grate	
Line 3	
Number of lines:	1
Waste capacity per line:	25.0 Mg/h
Thermal capacity per line:	87.0 MW
Steam output per line:	103.5 Mg/h
Steam pressure:	40 bar
Steam temperature:	400°C
Line 4	
Number of lines:	1
Waste capacity per line:	29.0 Mg/h
Thermal capacity per line:	90.0 MW
Steam output per line:	103.5 Mg/h
Steam pressure:	40 bar
Steam temperature:	400°C

Martin Waste-to-Energy Technology. Figure 44
Malmö (SE)

therefore does not increase the global CO_2 balance, similar to the combustion of biomass. The other half, however, is fossil in origin (plastics, paints, varnishes, etc.) and therefore impacts on the climate. Consequently, the technological groundwork must be laid for building highly efficient waste combustion plants and in order that the substitution effect of standard fuels can be maximized to the greatest extent possible. Far-reaching climate policies are basically inconceivable without these plants because their potential for reducing the greenhouse effect and as a source of alternative energies cannot be ignored. Development goals must serve to increase the efficiency of these plants while integrating them into the energy and material flow management systems of individual countries [35, 36].

Further development and optimization of existing technologies and concepts are needed due to international requirements in the field of thermal waste treatment using grate-based combustion systems. It has proven itself repeatedly that innovative technologies



Reverse-acting grate Vario

RDF

Number of lines:	1
Waste capacity per line:	13.4 Mg/h
Thermal capacity per line:	49.9 MW
Steam output per line:	61.2 Mg/h
Steam pressure:	56 bar
Steam temperature:	400°C

Martin Waste-to-Energy Technology. Figure 45
Pozzilli (IT)

Horizontal grate

1994

Number of lines:	4
Waste capacity per line:	30.0 Mg/h
Thermal capacity per line:	73.0 MW
Steam output per line:	77.0 Mg/h
Steam pressure:	43 bar
Steam temperature:	415°C

2007

Number of lines:	2
Waste capacity per line:	33.6 Mg/h
Thermal capacity per line:	93.3 MW
Steam output per line:	102.0 Mg/h
Steam pressure:	130 bar
Steam temperature:	420°C

Martin Waste-to-Energy Technology. Figure 46
Amsterdam (NL)

must first be developed and comprehensively investigated. In the future, MARTIN will continue to reliably ensure treatment of waste under ecological and economic constraints, using innovations and reliable process-engineering technology and taking international statutory requirements into account.

Bibliography

Primary Literature

1. MARTIN GmbH für Umwelt- und Energietechnik (2011) http://www.martingmbh.de
2. Koralewska R (2006) MARTIN Reverse-acting grate system – The challenge of high heating value fuels. In: Proceedings of the 14th NAWTEC, Tampa, 1–3 May 2006
3. Gohlke O, Busch M (2001) Reduction of combustion by-products in WTE plants: O_2 enrichment of underfire air in the Martin Syncom process. Chemosphere 42:545–550
4. Gohlke O, Koralewska R, Zellinger G, Takuma M, Kuranishi M, Yanagisawa Y (2006) Alternatives to ash melting and gasification. In: Proceedings of the 4th i-CIPEC, Kyoto, Japan, 26–29 Sept 2006
5. Gohlke O, Martin J (2007) Drivers for innovation in Waste-to-Energy technology. Waste Manage Res 25:214–219
6. Gleis M (2009) Reliability of new technologies of thermal waste treatment. In: 10th Assises des déchets, Atlantia la Baule, France, 21–22 Oct 2009
7. Gleis M (2010) Ungläubiges Kopfschütteln-Pyrolyse das einst viel gepriesene Abfallbehandlungsverfahren hat sich in Europa nicht durchgesetzt. Steht ihm nun eine Renaissance bevor? RECYCLING Magazine 07, pp 30–31
8. Tadishi O, Hiroyuki H, Kazuaki S (2006) Operation data of MSW gasification and melting plant. In: Proceedings of the 4th i-CIPEC, Kyoto, Japan, 26–29 Sept 2006
9. Whiting K, Schwager J (2006) Why are novel technologies, such as gasification, for MSW processing struggling to make an impact in Europe? In: Proceedings of the 4th i-CIPEC, Kyoto, Japan, 26–29 Sept 2006
10. Vehlow J (2009) Abfallverbrennung in Deutschland. Müllhandbuch digital, ESV Erich Schmidt Verlag, MuA Lfg. 2/09
11. International Energy Agency IEA (2010) IEA Biomass Agreement. Task X, Sub-task 6-Gasification of waste. http://www.ieabioenergy.com
12. Beckmann M, Scholz R, Wiese C, Davidovic M (1997) Optimization of gasification of waste materials in grate systems. In: International conference on incineration & thermal treatment technologies, San Francisco-Oakland Bay, 12–16 May 1997
13. Davidovic M (2007) Gasification and NO_x-reduction with a reverse-acting grate and a multistaged postcombustion chamber. Presentation at Clausthaler Umwelttechnik-Institut GmbH/CUTEC, Clausthal-Zellerfeld, Germany
14. Schreiner R, Jansen A (1997) Infrared cameras guide combustion control. Mod Power Syst 17(9):45–49
15. Meile E, Schreiner R (2002) Gezielte Prozessbeeinflussung durch Aufschalten einer Infrarotkamera am Beispiel der MVA Winterthur. Entsorgungspraxis 5:26–30
16. Zipser S, Gommlich A, Matthes J, Keller H, Fouda Ch, Schreiner R (2004) On the optimization of industrial combustion processes using infrared thermography. In: Proceedings of the 23rd IASTED international conference on modeling, identification and control, Grindelwald, pp 386–391
17. Keller H, Matthes J, Zipser S, Schreiner R, Gohlke O, Horn J, Schönecker H (2007) Kamerabasierte Feuerungsregelung bei stark schwankender Brennstoffzusammensetzung. VGB Powertech 3:85–92
18. Gohlke O (2009) Efficiency of energy recovery from municipal solid waste and the resultant effect on the GHG balance. Waste Manage Res 27:894–906
19. BREF of waste incineration (2005) Integrated pollution prevention and control. Reference document on the best available techniques for waste incineration (BREF), IPPC Bureau. http://eippcb.jrc.es
20. Gohlke O, Seitz A, Spliethoff H (2007) Innovative approaches to increase efficiency in EfW plants – Potential and limitations. ISWA, Amsterdam
21. Murer MJ, Spliethoff H, van Berlo MAJ, de Waal CMW, Gohlke O (2009) Comparison of energy efficiency indicators for EfW plants. Proceedings of the Sardinia 2009 symposium, Sardinia, Italy
22. Mennessier A (2008) Study of an innovative process for NO_x reduction in an energy-from-waste plant, BSc Technische Universität München, München
23. ANSYS Inc (2010) ANSYS FLUENT 12.0 Documentation
24. Wolf Ch (2005) Erstellung eines Modells der Verbrennung von Abfall auf Rostsystemen unter besonderer Berücksichtigung der Vermischung – Ein Beitrag zur Simulation von Abfallverbrennungsanlagen. Dissertation, Universität Duisburg-Essen, Duisburg
25. Martin U (2010) Beschreibung der Brennstoffumsetzung im Brennbett von Rostfeuerungen. Technische Universität München, München
26. Koralewska R (2005) Industrial-scale validation of a CFD simulation in conjunction with a fuel-bed model. Presentation at waste-to-energy research and technology council, Columbia University, New York
27. Koralewska R, Wolf Ch (2005) Industrial-scale validation of a CFD model in conjunction with a fuel-bed model for the thermal treatment of waste in grate-based combustion plants. VDI-Berichte 1888, 22. Deutscher Flammentag, VDI-Verlag, Düsseldorf, pp 613–619
28. Rossignoli P (2010) High-dust selective catalytic NO_x reduction at WTE plant in Brescia. In: Proceedings of the 2nd international conference on biomass and waste combustion, Oslo, Norway, 16–17 Feb 2010
29. Gohlke O, Busch M, Horn J, Takuma M, Kuranishi M, Yanagisawa Y (2003) New grate-based Waste-to-Energy system producing an inert ash granulate. Waste Management World May–June, pp 37–46

30. Martin J, Gohlke O, Tabaries F, Praud A, Yanagisawa Y, Takuma M (2005) Defining inert – a technological solution to minimize ecotoxicity. Waste Management World Sept/Oct, pp 70–73
31. Koralewska R (2009) SYNCOM-Plus: An optimized residue treatment process. In: Proceedings of the 17th NAWTEC, Chantilly, 18–20 May 2009
32. Schlumberger S (2010) Neue Technologien und Möglichkeiten der Behandlung von Rauchgasreinigungsrückständen im Sinne eines nachhaltigen Ressourcenmanagements. In: Proceedings of Bundesamt für Umwelt BAFU: Verbrennungsrückstände in der Schweiz, Bern
33. BSH Umweltservice AG (2011) http://www.bsh.ch
34. Schlumberger S (2005) Entwicklung und Optimierung eines Verfahrens zur selektiven Zinkrückgewinnung aus sauren Ascheextrakten der thermischen Abfallentsorgung. Dissertation, Technische Universität München, München
35. Fleck E (2006) A long-time flame: Waste-to-energy still goes strong in Europe. Waste Management World July–Aug, pp 107–116
36. Martin J (2010) Der Anlagenbau für Abfallverbrennungsanlagen – Strukturen und Märkte im Licht der Globalisierung. In: Thomé-Kozmiensky KJ, Versteyl A (eds) Planung und Umweltrecht, Band 4. TK, Neuruppin, Germany, pp 41–56

Books and Reviews

Chandler AJ, Eighmy TT, Hjelmar O, Vehlow J et al (1997) Municipal solid waste incinerator residues. Elsevier, Amsterdam

Ludwig CB et al (1973) Handbook of infrared radiation from combustion gases. NASA SP-3080, Washington, DC

Ortiz de Urbina G, Goumans JJJM (2003) Proceedings of the 5th International Conference on the environmental and technical implications of construction with alternative materials. WASCON, San Sebastian, Spain, 4–6 June 2003

Spliethoff H (2009) Power generation from solid fuels. Springer, Berlin

Mass Transit Science and Technology, Introduction

Gary L. Brosch
Center for Urban Transportation Research,
University of South Florida, Tampa, FL, USA

Mass transit, or public transportation, has and will continue to play a key function in the shaping of communities. Public transportation plays an enormous role in the economic and social well-being of individuals and societies. Some have said the study and understanding of public transportation is not "rocket science." Indeed, it is much more difficult, as it involves not just science, but also public policy and individual choices.

Mass Transit Science and Technology, Introduction of the *encyclopedia* will examine a wide range of public transportation alternatives from conventional buses and steel wheeled trains to state-of-the-art BRT (bus rapid transit) systems and magnetically levitated trains. It will examine specially designated high-occupancy vehicle lanes (HOT lanes) and the role of bicycle transport. Also examined will be public policy decisions concerning a wide range of issues from parking policy to land use decisions related to transit-oriented development.

Bus Rapid Transit (BRT) combines technological advances and public policy changes to provide a transit system that is faster than conventional buses and often less expensive and more flexible than light rail systems. Development of BRT systems is occurring worldwide as examined in ► Bus Rapid Transit: Worldwide History of Development, Key Systems and Policy Issues. Technological advances such as automated guidance and signal priority are of use only if institutional issues are addressed. Public policy decisions that play a crucial role in the implementation of such systems are addressed in ► Bus Rapid Transit, Institutional Issues Related to Implementation.

As Bus Rapid Transit systems have developed, so has the controversy and debate over choice between Bus Rapid Transit and conventional light rail (LRT) systems. Which is better in the long run, which is more costly, which will attract more riders, which is better for economic development have all been the questions in this ongoing debate. Two articles, such as ► Bus Rapid Versus Light Rail Transit: Service Quality, Economic, Environmental and Planning Aspects and ► Bus Rapid Transit and Light Rail Transit Systems: State of Discussion, provide useful insight into the issues involved. Included in these discussions is a review of the trends of the last 50 years in urban transportation, the differences between developed and developing countries, the characteristics of light rail systems and bus rapid transit systems, and direct comparisons of the alternative modes.

The use of public transportation is determined by a wide array of direct factors including its cost, convenience, travel time, and perceived safety. Ridership also is affected by external factors such as the cost of alternatives, congestion on roadways, and economic conditions. Public transportation is both affected by land use patterns and affects land use. Suburban versus urban developments clearly affect transit ridership potential. Similarly, transit centers clearly promote greater density. Balancing public policy goals for land use development and for use of mass transit can be facilitated through the use of transit-oriented development (TOD). ► Bus Versus Rail Implications for Transit-Oriented Development examines the differences and opportunities for transit-oriented development when comparing light rail systems and bus rapid transit systems. Urban transport systems clearly affect economic productivity and economic development. It is important to understand the differing implications of bus, bus rapid transit and rail systems on commercial and residential development surrounding transit stations.

Recent technological advances have enabled policy makers the opportunity to consider managing highway express lanes with high occupancy toll (HOT) lanes. Managing congestion and effective throughput of major expressways is a challenge. The use of high-occupancy vehicle lanes is gaining increased acceptance as is the increasing use of toll roads to finance new construction. Finding methods and technologies to combine these two offers the potential to increase the effectiveness of major roadways. Two articles, ► HOT Lanes/Value Pricing: Planning and Evaluation of MultiClass Service and ► High-Occupancy Vehicle and Toll Lanes provide an understanding of the technological advances and policy issues surrounding the potential of HOT lanes. Congestion pricing alternatives, equity issues, and future developments are included.

Light Rail Transit (LRT) is a staple of mass transit systems in both developed and developing countries. Its applications range from streetcars and tramways to modern automated systems. In spite of significant history and experience with light rail systems, understanding the conditions under which new or expanded systems are appropriate remains a challenge. ► Light Rail Transit, Systemic Viability provides important historical background on the development of LRT, a look at over 25 new systems developed in the last 30 years, as well as an analysis of the use of transit in cities with and without LRT.

As new light rail systems have been implemented and existing systems expanded, there have been increasing concerns related to their operation and interaction with pedestrians in urban areas. Light rail is most effective when pedestrian interface is easiest. Yet, this easy access leads directly to greater safety concerns. Technological and policy innovations such as regulatory and warning devices, delineation marking and positive control devices offer potential to reduce pedestrian injuries. ► Light Rail Transit in the US and Abroad, Examination of History and Innovations offers a comprehensive examination of the latest innovations in pedestrian/LRT safety.

Traveling on a cushion of air in high-speed trains is a dream that technological innovation has brought to reality in the form of magnetically levitated trains, both high speed and low speed. ► MAGLEV Technology Development provides the reader with a comprehensive review of magnetically levitated vehicles. The physics, engineering, and costs of magnetic levitation are covered as well as the history and state of the art of its development.

High-Speed Rail systems developed in the last 50 years continue to be expanded and implemented in countries all over the world playing a significant role changing public transport opportunities. Economic changes and technological advances have led to the increasing popularity of high-speed trains, particularly in Europe and Asia. ► High Speed Rail, Technology Development of examines technological developments of high-speed rail as well as the resultant challenges of implementation.

Light rail systems can be effective in moving larger numbers of people from a limited number of origins and destinations. Whereas, moving smaller numbers of people from greater numbers of origins to greater numbers of destinations calls for a smaller-scale system. Personal Rapid Transit (PRT) systems attempt to provide such a smaller-scale system. Sometimes referred to as automated people movers, these systems offer an intriguing concept of quickly moving small groups of people from one place to another. ► Personal Rapid Transit and Its Development looks at the history of PRT in several countries and the current state of PRT development. Also provided is a discussion of the sustainability/energy issues and planning/architectural issues related to the future of personal rapid transit systems.

At the other end of the speed spectrum from high-speed rail is the integration of bicycle transport with mass transportation. Often overlooked, bicycles are an important element of a comprehensive transit system. Finding techniques to successfully integrate the use of bicycles with mass transit bus and rail systems can significantly increase the overall effectiveness of transit systems. A series of case studies and best practices is found in ► Bicycle Integration with Public Transport.

Transit-oriented development, sometimes referred to as "smart growth" seeks to integrate public transportation investments and land use near transit stations to improve the sustainability and effectiveness of both. ► Transit-Oriented Development and Land Use provides a comprehensive look at such developments in countries ranging from Sweden, to Columbia and to the USA. Successful examples of the transportation and environmental benefits of transit-oriented development are presented. Also presented are the challenges, the financing, and the future of transit-oriented development.

► Advanced Public Transport Systems, Simulation-based Evaluation offers the state of the art in computer simulation of transit operations. The use of these simulation techniques is valuable in areas such as fleet management, traveler information systems, electronic fare payments, and transportation demand management systems. Examples are provided of both microscopic transit simulation and mesoscopic transit simulation.

Medical Device Batteries

Michael J. Root
Cardiology, Rhythm and Vascular Research and
Development, Boston Scientific Corp., St. Paul,
MN, USA

Article Outline

Glossary

Anode The negative electrode of a discharging cell or battery.

Cathode The positive electrode of a discharging cell or battery.

Electrolyte A solution or material that completes the electrical circuit in a cell by way of ionic conduction.

Hermetic seal A way to seal implantable medical device batteries that is impermeable to fluids and usually includes a terminal feed through that is sealed in glass.

Implantable defibrillator An implanted medical device that functions as a pacemaker and is also capable of delivering high energy shocks to the heart to treat ventricular tachycardia (abnormally fast heart rate) and fibrillation.

Neurostimulation or neuromodulation Electrical stimulation of nerves to modify nerve activity.

Pacemaker An implanted medical device that delivers low level electrical stimulation to the heart for treatment of bradycardia (abnormally slow heart rate).

Primary cell A cell that is intended to be discharged only.

Secondary or rechargeable cell A cell that can be charged following depletion using an external electrical energy source.

Definition of the Subject

Medical device batteries serve an important role in modern health care. They power the devices that allow patients to function more normally by managing and improving their health or even survive life threatening disease conditions.

There are several ways to classify medical devices. Some provide therapeutic functions while others are used for diagnostic purposes. Some provide both functions. For therapeutic devices, devices may be further described as life sustaining or life enhancing (improves the quality of life).

Life sustaining devices provide therapy that keeps a patient alive. An implantable defibrillator is an example of a life sustaining device. Certain cardiac patients experience tachycardia (a very rapid heart beat) or ventricular fibrillation. Either condition limits the ability of the heart to effectively pump blood. Death can result if this type of arrhythmia is sustained and a normal heart rhythm is not restored by a defibrillator.

Life enhancing devices might treat conditions, like severe, chronic lower back pain, that do not threaten a patient's life, but prevent normal function and reduce their quality of life.

Improving the well-being of the patient is the most important use of wearable or implantable medical devices, of course. Medical devices can provide effective treatment or diagnostic information. Although it is not a necessary consideration, treating or monitoring a patient's condition remotely – that is, outside of the hospital or clinic – with such devices can also impact energy sustainability. Patients do not need to be transported to the hospital or clinic as frequently for treatment or monitoring.

In this entry are discussed a few of the specialized batteries for medical devices that are portable or wearable (carried with the patient, like hearing aids), or implantable (surgically placed inside the body as with neurostimulation pain management devices). There is a focus on the batteries designed for a few of the more common applications – implantable cardiac rhythm management (cardiac pacemakers and defibrillators), pain management, and hearing loss devices.

Introduction

Medical Devices that Use Batteries

There are a substantial number of wearable and implantable medical devices powered by batteries. These include devices for cardiac rhythm management (pacemakers, defibrillators, and heart failure devices), hearing loss, bone growth and fusion, drug delivery for therapy or pain relief, nerve stimulation for pain management, urinary incompetence and nervous system disorders, vision, diagnostic measurements and monitoring, and mechanical heart pumps.

This entry will be limited to three major categories of medical devices – cardiac rhythm, neurostimulation, and hearing devices. The first two are devices that apply electrical stimuli to muscle tissues or nerves and the last involves sound amplification. The battery systems used by these devices are used in other devices, as well. Some of these will be noted in the sections describing the batteries.

General Design Considerations for Medical Device Battery Performance

There are several important features that battery developers must consider when designing batteries for medical devices. Many of these are also important for most other battery types, as well.

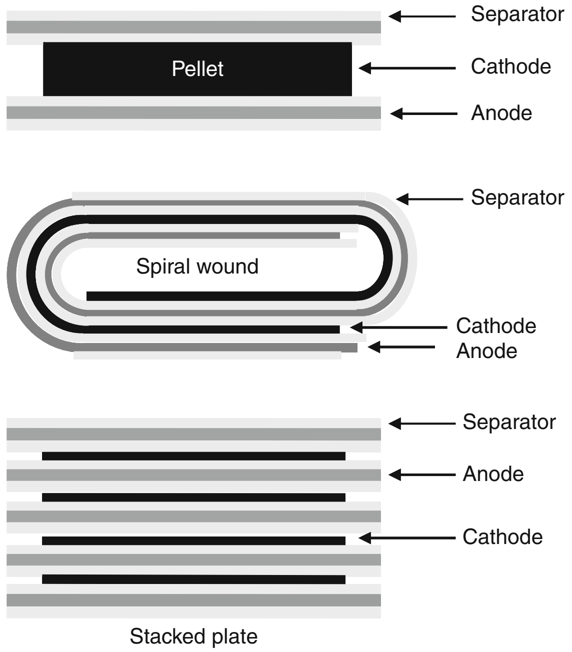
Medical device batteries are fundamentally the same as any other battery designed for consumer electronics, military, or aerospace applications. All require the same three components to be able to function as an electrochemical power source – a negative electrode (or anode) material to supply electrons, a positive electrode (or cathode) material that takes up electrons, and an electrolyte that completes the electrical circuit through ionic conduction. The other components in a cell are necessary to make the cell perform efficiently, minimize its size, and make it safe and reliable. These components include one or more separators that are electrically insulating to prevent direct contact between the anode and cathode but allow ions to pass through, current collectors to convey electrons to or from the electrodes and various insulators to prevent short circuits.

Medical devices, especially implantable devices, often use batteries that are custom designed and built specifically for that device. Their sizes and shapes may not be standard. The electrode assemblies may consist of pellet or slug electrodes, spirally wound electrode foils, or stacked electrode plates (Fig. 1).

Batteries for implantable medical devices are hermetically sealed. Hermetic seals have long been used for certain cell types, like lithium-sulfur dioxide and lithium-thionyl chloride, where long shelf life is important, or exposure to corrosive and toxic materials could result if the cell leaks.

A feedthrough terminal is used in hermetically sealed cells as a way to connect to one of the battery electrodes to the external device circuit. A glass to metal seal insulates the terminal from the cell case. The glass is specially formulated to create a tight seal against the battery case and the terminal and resist attack by the cell electrolyte. Glass materials like TA-23 and Cabal-12 are common.

A number of different chemistries have been used for medical device batteries. For example, over 20 years ago there were at least 14 battery chemistries developed and used in implantable cardiac rhythm devices for treating cardiac arrhythmias (Fig. 2). Only a few of these survived and are in use today.



Medical Device Batteries. Figure 1
Electrode assembly designs for prismatic medical device batteries

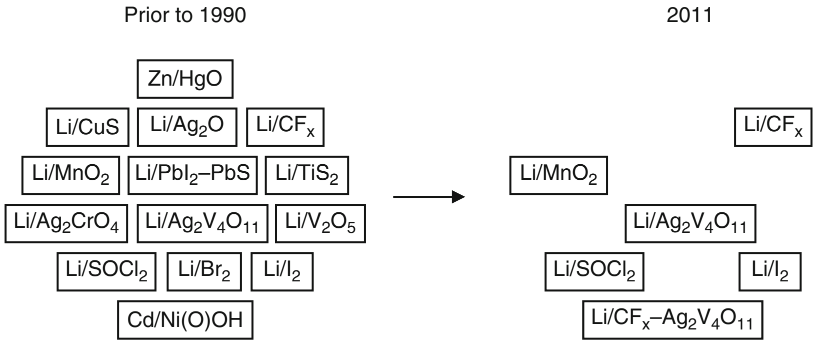
Electricity as Medicine

Physiological effects of electricity were known by the time Italian scientist Alessandro Volta first invented the battery in 1800. The introduction of the battery meant that a sustained source of electricity was available for the first time. Prior to that, electrostatic sources of electricity and storage devices, like the Leyden jar, were used. Such devices provided only momentary electrical current.

The electrochemical cell, or battery, enabled advances in many areas of research and development involving electricity during the nineteenth century including electrochemistry, electrophysiology, and the use of electricity as medical therapy.

The work of Guillaume-Benjamin-Amand Duchenne (de Boulogne), a French neurologist, was foundational to the advancement of electrophysiology. German neurologist Robert Remak was an early proponent of electrotherapy to treat diseases of the nervous system.

J. Althaus wrote in his 1860 book “A Treatise on Medical Electricity, Theoretical and Practical,” [1] “We know that, whatever may be the properties of the nerves, they can be called into action by galvanism. But the effects are widely different according to the form of electricity that is used; again, the quality and intensity of electricity are both of great importance; not less so the mode in which it is transmitted to the human body, and the length of time during which its action is kept up. In fact, we are able, by merely varying modes of applying electricity, to arouse or to kill the vital power of the nerves, and to diminish or to increase



Medical Device Batteries. Figure 2
Types of battery chemistries used for cardiac rhythm devices prior to 1990 and today

their properties. Hence electricity can only be expected to be of service in the treatment of disease, if we are guided in its use by an exact knowledge of the physiological effects which it will invariably produce."

In 1873, Herbert Tibbits wrote, "There will be found here no new ground opened out, but only an earnest endeavor to sift the wheat of our existing knowledge from the chaff, and to make the reader as much at home with his electrical as with his other medical instruments; and further to lead him to estimate electricity at its fair and proved value in therapeutics, as an agent, not to be indiscriminately advocated as a panacea, nor, on the other hand, neglected by the inexperienced, but in appropriate cases to be regarded as one of the most powerful and serviceable weapons with which we can combat disease"[2].

Many claims of treatments using electricity were false, though. H. Lewis Jones in 1892 cautioned the medical community, [3] "One thing is certain, that without a thorough grounding in the physical part of the subject, no satisfactory advances can be made in a field of therapeutics which is at present almost entirely neglected by medical men. A great deal of the quackery which surrounds and discredits medical electricity, is due to the indifference and contemptuous attitude of the medical profession, and we have only ourselves to blame if the public insist on seeking elsewhere for treatment which is refused to them by their medical advisors."

Perhaps, the most prominent and successful use of electrical stimuli in medicine today is for cardiac rhythm management used to treat certain cardiac disease conditions. The American Heart Association reports that nearly 37% of the US population 20 years and older have some form of heart disease. The Heart Rhythm Foundation estimates 325,000 deaths each year in the USA from sudden cardiac arrest making it a leading cause of death.

There are three implantable devices used today to treat cardiac arrhythmias – the pacemaker, the implantable cardioverter defibrillator (ICD), and the cardiac resynchronization therapy (CRT) devices for heart failure patients.

Implantable Cardiac Pacemakers

Blood flow throughout the body occurs because the pumping action of the heart maintains the arterial blood pressure at a higher level than the venous blood pressure. Each heart beat involves the coordinated contraction of cardiac muscle cells as the result of the initiation and propagation of an electrical impulse [4].

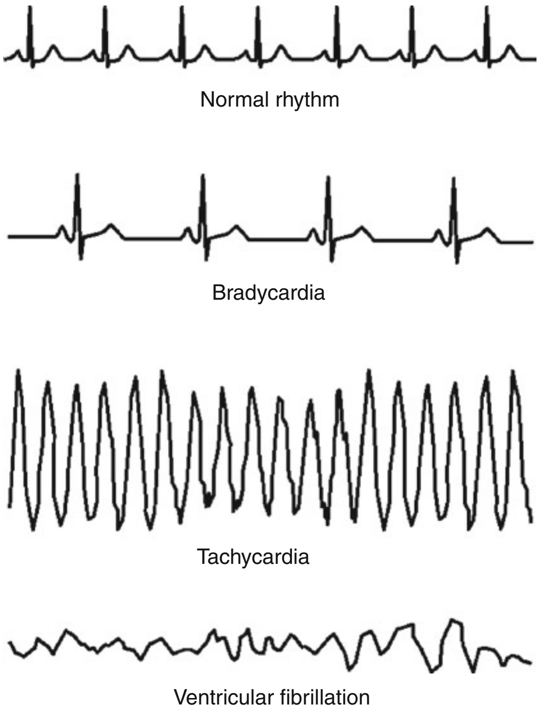
Rather than contracting from nerve stimulation like skeletal muscles, cardiac muscle or myocardial cells respond to an electrical signal created by ionic (K^+, Na^+, and Ca^{2+}) concentration differences across their cell membranes. At rest, this potential, called an action potential, is about −90 mV. When the cell is activated, it rapidly depolarizes by redistributing the ionic concentrations across the cell membrane and contraction follows. Specialized muscle cells in the sinoatrial (SA) node of the heart start this process by spontaneously depolarizing once it reaches the critical firing level (CFL) of about −55 mV. The SA node is the primary natural pacemaker in normally functioning hearts. The signal is then quickly conveyed throughout the rest of the heart by way of several conduction pathways [4].

There are a number of conditions that cause the heart to beat abnormally. Electrocardiograms measure changes in potential that occur in the heart. An electrocardiogram of a normal heart rhythm is shown in Fig. 3. Electrocardiograms showing a normal rhythm, bradycardia, tachycardia, and ventricular fibrillation. The beats are regular and the time intervals between them are typical.

Bradycardia is a condition wherein the heart beats too slowly or skips beats altogether (for example, see Fig. 3). This can occur if the electrical impulses through the heart are slowed, too few impulses are transmitted, or they are blocked altogether [4].

Implantable cardiac pacemakers are electronic devices that treat bradycardia. External pacemakers were developed in the early 1950s. These were large devices that were not portable. Introduction of the transistor in the mid-1950s meant that smaller devices could be built. Earl Bakken and coworkers later developed a smaller external pacemaker that could be carried by the patient [5]. The first cardiac pacemakers to be implanted in humans lasted hours to days [6]. In 1960, Wilson Greatbatch developed the first implantable cardiac pacemaker that lasted for longer than a few days [7].

Pacemakers today can pace in one or both chambers (right atrium and right ventricle) on the right side of the heart. The power demand on a pacemaker battery



Normal rhythm

Bradycardia

Tachycardia

Ventricular fibrillation

Medical Device Batteries. Figure 3
Electrocardiograms showing a normal rhythm, bradycardia, tachycardia, and ventricular fibrillation

of any individual device varies based on device programming, the number of chambers paced, and the patient's cardiac condition. Cardiac pacing therapy is a relatively low power operation – generally between 20 and 100 μW on average. The amount of electrical energy used to stimulate the heart can be adjusted to meet the needs of the patient. The first pacemakers where asynchronous – meaning they paced at a single rate. Devices today adapt to the patient's activity level to pace more rapidly or more slowly as needed. Additional low power demands on these devices include sensing to detect a natural heart beat and other monitoring features.

Although requiring only low levels of power, it is the frequent application of pacing therapy, as well as the continuous background device operations, that have the most significant impacts on device longevity. Implantable cardiac pacemakers are expected to last up to 10 years or so, depending on the therapy required by the patient; so a battery that has a high energy density is important.

The first devices to be implanted in humans used either rechargeable nickel-cadmium (NiCd) cells or alkaline zinc-mercuric oxide (Zn/HgO) cells [8].

The development of primary lithium batteries for implantable medical devices was a big advance that enabled devices to operate more reliably and longer. Lithium is the lightest metal and has the most negative reduction potential. When combined with any number of positive electrode materials, the result is cells with high energy densities compared to aqueous cells. Most lithium cells have an initial open circuit voltage between 1.8 and 3.9 V, compared to 1.2–1.6 V for most aqueous cells.

A number of lithium cell technologies were developed in the 1970s [9] and some were used clinically, [10] including lithium-copper sulfide (Li/CuS), lithium-silver chromate (Li/Ag_2CrO_4), lithium-thionyl chloride (Li/$SOCl_2$), lithium-lead sulfide-lead iodide (Li/PbS-PbI_2), lithium-titanium disulfide-sulfur (Li/TiS_2-S), lithium-bromine (Li/Br_2), lithium-manganese dioxide (Li/MnO_2), and the lithium-iodine (Li/I_2) cells that are used in most pacemakers today.

Other power sources were developed, as well. Notable were radioactive power sources, principally based on plutonium-238, [11] that were used in the 1970s. Regulations pertaining to the distribution of radioactive devices limited their acceptance.

Implantable Cardiac Pacemaker Batteries

Lithium-Iodine Lithium-iodine (Li/I_2) was proposed as a power source for implantable cardiac pacemakers in 1971 [12]. The first pacemaker run by a Li/I_2 cell was implanted in 1972 [6]. These cells were originally developed as more reliable and longer lived alternative to the zinc-mercuric oxide cells (see below) used in implantable cardiac pacemakers since they were introduced in 1960 and on into the mid-1970s. However, Li/I_2 cells have been the dominant power source for implantable cardiac pacemakers for more than 30 years.

The negative electrode reaction is:

$$\mathrm{Li} \rightarrow \mathrm{Li}^{+} + \mathrm{e}^{-}$$

and the positive electrode reaction is:

$$\mathrm{I}_2 + 2\mathrm{e}^{-} \rightarrow 2\mathrm{I}^{-}$$

Iodine is stabilized by mixing it with a polymeric pyridine – poly-2-vinylpyridine (abbreviated P2VP or PVP) – to form an I_2-P2VP charge transfer complex.

The overall cell reaction is:

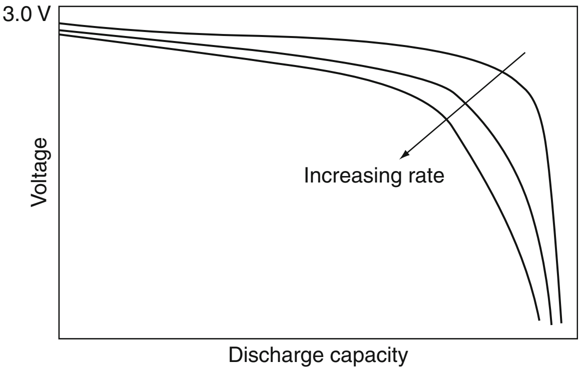
$$2Li + I_2 \rightarrow 2LiI$$

When a Li/I_2 cell is built and the I_2-P2VP mixture is added, it reacts directly with Li to form in situ a layer of solid LiI between negative and positive electrode materials. This layer serves to protect the Li from further reaction with the I_2. LiI also functions both as an electrolyte and a separator.

The movement of Li^+ is rather slow in solid LiI with conductivities less than $10^{-6}\ \Omega^{-1}\ cm^{-1}$ [13]. This limits Li/I_2 cells to low rate applications, like implantable cardiac pacemakers.

The open circuit potential of the Li/I_2 cell is about 2.8 V. As the cell discharges, the thickness of the LiI electrolyte layer increases. This increases the internal resistance of the cell, so the discharge voltage decreases somewhat as the cell becomes depleted. The faster the discharge rate, the more sloped the discharge voltage becomes (Fig. 4).

The gradual voltage and resistance changes during discharge present the means to predict the extent of cell depletion, thereby allowing an accurate prediction of pacemaker longevity. This advanced warning allows sufficient time for the physician and the patient to schedule device replacement well ahead of complete battery depletion and the subsequent inability of the pacemaker to deliver therapy.



Medical Device Batteries. Figure 4
Discharge voltage of Li/I_2 cells. The voltage is lower for higher discharge loads

A cross section of a Li/I_2 prismatic cardiac pacemaker cell is shown in Fig. 5. In this design, a lithium sheet is placed between two I_2-P2VP electrodes. Energy densities for Li/I_2 cells are 210–280 Wh kg^{-1} and 810–1,030 Wh dm^{-3} [14].

Other devices that use Li/I_2 cells are implantable bone growth stimulators [15] and an implantable electromechanical hearing system [16].

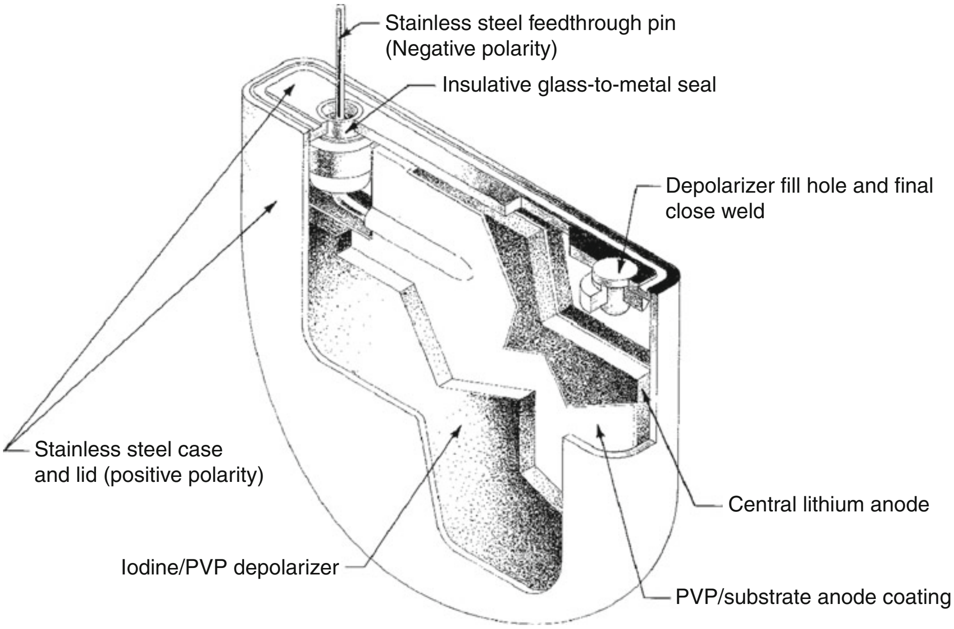
Lithium-Carbon Monofluoride The most common positive electrode materials that have been used in all types of batteries, including those designed for medical devices (I_2 and O_2 notwithstanding), are usually metal oxides, such as silver vanadium oxide ($Ag_2V_4O_{11}$), manganese dioxide (MnO_2), mercuric oxide (HgO), and silver oxide (Ag_2O), or nonmetal oxides, like thionyl chloride ($SOCl_2$). Oxide compounds have a high energy density and can be chemically stable.

A cell pairing lithium (Li) with elemental fluorine (F_2) would have a theoretical open circuit voltage of 5.9 V. It is not practical to build a cell using elemental fluorine, which is a gas at room temperature and highly reactive, but fluoride compounds that retain some of the high energy density could be more realistic positive electrode materials [17].

Carbon monofluoride (CF_x) is one such fluoride compound that today is used as a positive electrode material in Li batteries for a number of different applications. For example, they are used in certain types of heart failure devices – implantable cardiac resynchronization therapy pacemakers (CRT-P). CRT-P devices can pace the right atrium and right ventricle, but they are also capable of pacing the left ventricle. Pacing three chambers requires more power than a Li/I_2 cell can deliver, so a different battery type is needed. Li/CF_x cells were developed in response to the increased power required by CRT-P devices. Vagal nerve stimulator devices also use a Li/CF_x cells.

Li/CF_x cells were first launched/introduced commercially in the mid-1970s and are available as coin cells and spiral wound cylindrical cells in a variety of shapes and sizes.

CF_x used in batteries is synthesized by the direct reaction of fluorine gas with a carbon starting material, such as petroleum coke, at high temperatures – usually between 350°C and 600°C depending on the type of carbon starting material and the fluorination level.


Stainless steel feedthrough pin (Negative polarity)
Insulative glass-to-metal seal
Depolarizer fill hole and final close weld
Stainless steel case and lid (positive polarity)
Central lithium anode
Iodine/PVP depolarizer
PVP/substrate anode coating

Medical Device Batteries. Figure 5
Cross section of a Li/I_2 prismatic cardiac pacemaker cell. © Greatbatch, Inc (Reprinted with permission)

The value of x in CF_x used in batteries, where x is the average number of fluorine atoms per carbon atom throughout the CF_x material, is usually between 0.9 and 1.2.

Fluorine is a highly corrosive and hazardous gas and so is difficult to handle safely. As a result there are only a few companies that have the ability or desire to manufacture CF_x in large quantities.

The CF_x material is mixed with a conductive carbon, like acetylene black or graphite, and a binder to maintain the mechanical integrity of the electrode. Typical electrolyte solutions are comprised of lithium tetrafluoroborate ($LiBF_4$) dissolved in one or more organic solvents, generally propylene carbonate (PC) or gamma-butyrolactone (GBL) with 1, 2-dimethoxyethane (DME).

The overall cell reaction is:

$$x\text{Li} + \text{CF}_x \rightarrow \text{C} + x\text{LiF}$$

Lithium fluoride (LiF) precipitates as a reaction product and it can limit the utilization of the CF_x.

The nominal open circuit voltage for Li/CF_x is 3.0 V, but the discharge voltage is closer to 2.8 V. The internal resistance of the Li/CF_x cell decreases during the early stages of discharge and remains low throughout discharge (Fig. 6), [18] attributed to the formation of conductive carbon as a reduction product of CF_x.

Energy densities can range between 250–590 Wh kg^{-1} and 635–1,050 Wh dm^{-3} depending on the battery size [19].

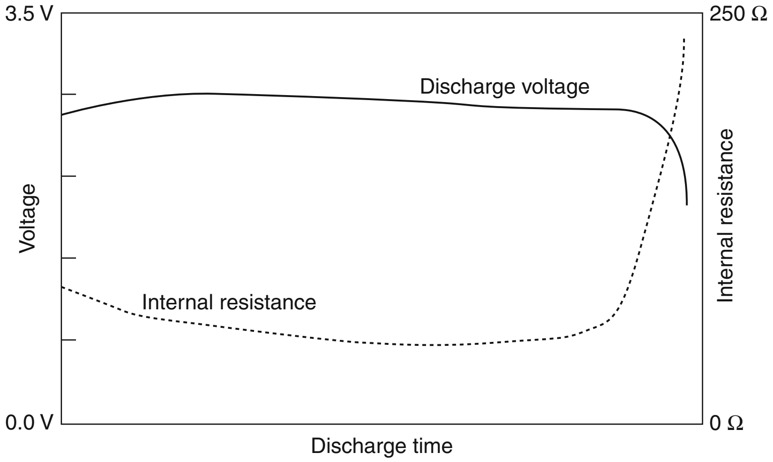
CF_x is also now combined with silver vanadium oxide designed for pacemakers and implantable defibrillators (see below).

Lithium Thionyl Chloride Most positive electrode materials are solids. However, among the first lithium cell types to be developed used an inorganic liquid, $SOCl_2$, as the positive electrode material.

Lithium-thionyl chloride ($Li/SOCl_2$) cells have seen a number of uses, including remote monitoring (such as residential water meters), various OEM (original equipment manufacturer) electronic devices, military, aerospace, and down-hole oil well monitoring applications. Medical device uses have included implantable heart monitors, drug infusion pumps, and some of the earliest implantable cardiac pacemakers [9, 10].

The positive electrode reaction is:

$$2\text{SOCl}_2 + 4e^- \rightarrow \text{S} + \text{SO}_2 + 4\text{Cl}^-$$



Medical Device Batteries. Figure 6
Discharge voltage and internal resistance for a Li/CF_x coin cell

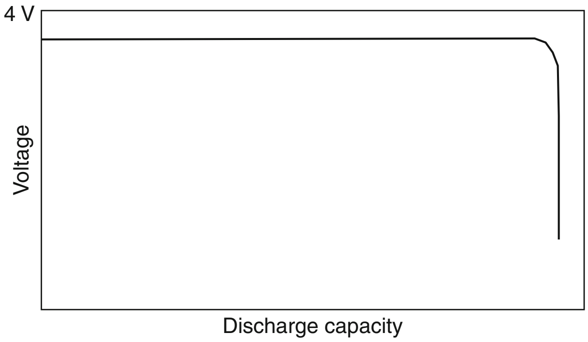
with the overall cell reaction:

$$4Li + 2SOCl_2 \rightarrow S + SO_2 + 4LiCl$$

Thionyl chloride serves as both positive electrode material and the electrolyte solvent. The electrolyte salt typically is lithium aluminum chloride ($LiAlCl_4$) or, sometimes, lithium gallium chloride ($LiGaCl_4$).

The design of the Li/$SOCl_2$ cell puts $SOCl_2$ in direct contact with the Li anode material. If the direct reaction between Li and $SOCl_2$ continues, the active electrode materials would produce no useful electrical energy. The reaction between Li and $SOCl_2$ is self-limiting, though. Similar to other lithium cell systems, reaction products form a passive layer on the Li that inhibits further reactions from occurring. The passive layer must be both electrically insulating yet ionically conductive for the battery to discharge efficiently.

In systems with solid cathodes, the active positive electrode material is usually mixed with a conductive carbon that conveys electrons from the current collector to the electrode active material. In this case (and similar to the zinc-air cell discussed below), liquid $SOCl_2$ is the active material and needs to come into contact with an electrode site (cathode) where it can take up electrons as the cell discharges. The cathode in Li/$SOCl_2$ cells is a porous carbon, such as acetylene black, and a PTFE binder. An aluminum screen can be used as a current collector to electrically connect the cathode and the positive cell terminal.



Medical Device Batteries. Figure 7
Discharge voltage for a Li/$SOCl_2$ cell

The lithium chloride (LiCl) and sulfur (S) that form precipitate and build up at the cathode. The cell capacity can be limited by these products if the cathode becomes blocked. Additionally, sulfur dioxide (SO_2) gas forms as a reaction product.

The Li/$SOCl_2$ cell has one of the higher open circuit potentials, 3.65 V, of any primary lithium cell. The voltage is quite flat throughout discharge (Fig. 7). The energy density is about 380 Wh kg^{-1} and 715 Wh dm^{-3}.[14] Li/$SOCl_2$ cells are produced in a wide variety of shapes and sizes for both high and low power applications, including prismatic and spiral-wound cylindrical cells used in pacemakers implanted during the 1970s.

Zinc-Mercuric Oxide Although no longer used, zinc-mercuric oxide (Zn/HgO) cells were the power source of choice for the first commercially viable implantable cardiac pacemakers. More than 3 million Zn/HgO cells were implanted in the 16 years from when the first successful cardiac pacemaker was implanted in 1960 and 1976 [20]. They helped many bradycardia patients until they were eventually supplanted by lithium cells, particularly Li/I$_2$, so a brief description of Zn/HgO technology is included here. Zn/HgO cells were also once used in wearable hearing aids.

The use of mercuric oxide as a positive electrode material for Zn/HgO cells was patented by British inventor Charles L. Clarke in 1884 [21], and also by others like Danish scientist Johannes N. Brønsted [22] in the early 1900s. A practical Zn/HgO cell was developed in the 1940s by Samuel Ruben with manufacturing support from the Mallory Battery Co [23]. The Zn/HgO cell was designed to replace the Leclanché zinc-manganese dioxide (Zn/MnO$_2$) cells used in military communications equipment during World War II. The Zn/HgO cell had better storage life and performance compared to Leclanché cells, particularly in the high temperature and high humidity conditions of the South Pacific [24].

Assorted Zn/HgO button and cylindrical cell sizes were produced after the war for military and space applications, as well as a number of consumer applications, including calculators, watches, and cameras. Some of the first implantable cardiac pacemakers that were available commercially were powered by Zn/HgO batteries and they became a popular cell for use in wearable hearing aids.

The Zn negative electrode material, or anode, and electrolyte solution are similar to other primary alkaline battery types, like zinc-air and zinc-silver oxide (Zn/Ag$_2$O). Zinc powder is mixed with a gelling agent like polyacrylic acid and a KOH-ZnO-H$_2$O electrolyte.

The positive electrode reaction is:

$$HgO + H_2O + 2e^- \rightarrow Hg + 2OH^-$$

The HgO is usually mixed with a conductive carbon powder, like graphite, to improve the electrical conductivity of the cathode. Mercuric oxide forms Hg as it is reduced. Mercury metal is highly conductive, and so it does lower the internal resistance of the cell. It is also a liquid at normal operating conditions, and it tends to coalesce into droplets and pool in a way that can reduce performance. Additives to the cathode like manganese oxides or silver can minimize Hg pooling.

The overall cell reaction may be written:

$$Zn + HgO + 2KOH + H_2O \rightarrow K_2Zn(OH)_4 + Hg$$

or

$$Zn + HgO \rightarrow ZnO + Hg$$

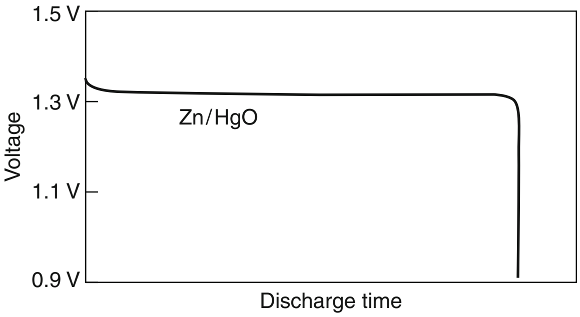
A cross section of a Zn/HgO button cell is similar to that of a Zn/Ag$_2$O cell and is shown in Fig. 16.

The initial open circuit potential of Zn/HgO cells is about 1.35 V, but can be between 1.40 V and 1.55 V if MnO$_2$ is added to the cathode. The voltage remains rather constant throughout discharge (Fig. 8).

Energy densities for Zn/HgO button cells are about 100 Wh kg^{-1} and 470 Wh dm^{-3} [14].

Mercuric oxide cells are not readily available any longer. Government regulatory agencies throughout the world forced these cells from the market because of environmental concerns.

For wearable hearing aids, other cells are available, including Zn/Ag$_2$O and zinc-air discussed below. Lithium-iodine cells replaced Zn/HgO batteries for implantable cardiac pacemakers in the starting, in the early 1970s. The Zn/HgO cells tended to generate gas which is difficult to manage in an implantable device. Further, actual device longevity – generally in the range of 18–36 months – was lower than the expected 5 or more years [20]. Early battery depletion was usually the cited cause; however, some claimed device malfunction was actually responsible [20].



Medical Device Batteries. Figure 8
Discharge voltage behavior for a Zn/HgO cell

Another concern was the voltage characteristics of Zn/HgO cells during use. The end of battery life is difficult to predict because the discharge voltage is flat through battery life then quite rapidly decreases when the battery becomes completely depleted (Fig. 8). The lithium cells developed to take the place of Zn/HgO cells were more reliable, longer lived, and the end of battery life is more easily anticipated.

Implantable Cardioverter Defibrillators and Heart Failure Devices

The implantable cardioverter defibrillator (ICD) is a cardiac pacemaker. It can be used to pace one or both chambers on the right side of the heart. It has an additional feature, though. ICDs can also impart powerful shocks to the heart if it is beating too fast (tachycardia) or goes into ventricular fibrillation. Either condition means that blood cannot be pumped very efficiently, if at all. A number of major clinical studies were done that identified various categories of heart patients who could benefit from the therapies delivered by devices like the ICD.

The cardiac resynchronization therapy defibrillator device (CRT-D) provides the same pacing and defibrillation functions as an ICD, but can also pace the left ventricle for heart failure patients.

The need to deliver energetic shocks to the heart within seconds presents a challenge for battery designers. The battery must provide years of operation for the constant or frequent low power demands of pacing, sensing, and other device functions in the tens of mW range and also the infrequent, but high power, pulses to shock the heart. To deliver a shock, the battery must charge high voltage electrolytic capacitors generally somewhere between 600 and 800 V. The shock is delivered by discharging the capacitors into the heart tissue. Rapidly charging the capacitors requires power on the order of watts. Sometimes multiple shocks are required to get the heart back to a normal rhythm. The balance between energy density for longevity and rate capability for rapidly charging the high voltage capacitors requires careful selection of the chemistry as well as the mechanical design of the battery.

The first clinical studies of ICDs implanted in humans started in 1980 [6] following years of development by Mirowski and coworkers [25]. The first commercially available device was released in 1985. Since then, large clinical studies identified new patient populations that could benefit from ICDs and CRT-Ds.

The first commercial devices were up to 160 cm^3 in volume and needed to be implanted in the abdomen because of their size. Device features were limited and nonprogrammable. Devices typically lasted 2 years before replacement was necessary.

Devices today are about one-fifth the size of the first ICDs. The device industry continued to make improvements by developing smaller devices and while adding more features, such as multichamber pacing, monitoring patient cardiac performance with recorded electrocardiograms, and remote monitoring using radio frequency (RF) telemetry, while providing greater longevity. Devices now are small enough to be implanted in the pectoral area of the chest and can last for 5 or more years.

Improvements in device electronics partially helped to improve device longevities, but new battery systems were also needed. The earliest ICDs were powered by lithium-vanadium pentoxide (Li/V_2O_5) cells [6]. Longevities were insufficient using these cells, so a different battery chemistry, lithium-silver vanadium oxide (Li/$Ag_2V_4O_{11}$ or Li/SVO) was developed [26]. Li/$Ag_2V_4O_{11}$ cell was the cell of choice for ICDs and CRT-D devices since 1985 until recently when other cell chemistries have become more prevalent.

Implantable Cardioverter Defibrillator Batteries

Lithium-Silver Vanadium Oxide The earliest implantable defibrillators used lithium-vanadium oxide (Li/V_2O_5) cells. The chemical stability of this type of cell was unsatisfactory, and they were soon replaced with lithium-silver vanadium oxide (Li/$Ag_2V_4O_{11}$ or Li/SVO) cells. Until only the last few years, Li/$Ag_2V_4O_{11}$ cells were by far the most common cell system used in implantable defibrillators.

Silver vanadium oxide is prepared by the high temperature reaction of V_2O_5 with a silver salt like $AgNO_3$ by a decomposition pathway [27]:

$$2V_2O_5 + 2AgNO_3 \rightarrow Ag_2V_4O_{11} + \text{nitrogen oxides}$$

or with Ag_2O through a combination mechanism [28]:

$$2V_2O_5 + Ag_2O \rightarrow Ag_2V_4O_{11}$$

Reduction to about 1.5 V vs. Li involves up to 7 electrons per mole of $Ag_2V_4O_{11}$:

$$Ag_2V_4O_{11} + 7e^- \rightarrow Ag_2V_4O_{11}{}^{7-}$$

though utilization is less under realistic use conditions.

The $Li/Ag_2V_4O_{11}$ cell discharge reaction occurs in several steps involving sequential reduction of Ag(I) to Ag(O) and V(V) to V(IV) [29]:

$$2Li + Ag^I{}_2V^V{}_4O_{11} \rightarrow Li_2Ag^0{}_2V^V{}_4O_{11}$$

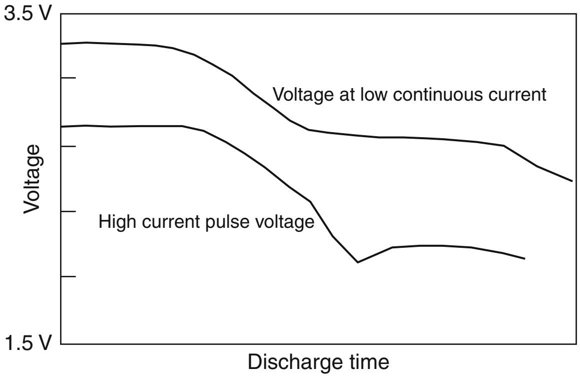
$$4Li + Ag^I{}_2V^V{}_4O_{11} \rightarrow Li_6Ag^0{}_2V^{IV}{}_4O_{11}$$

Some V(III) may also form, especially if the discharge proceeds beyond $Li_6Ag_2V_4O_{11}$.

Silver(I) is reduced to silver(0) in the first discharge step. Silver metal has a high conductivity, so the resistance of the $Ag_2V_4O_{11}$ positive electrode material decreases rapidly during the early stages of discharge.

The different discharge steps occur at different cell potentials (Fig. 9) that result in a series of voltage steps. The distinct voltage levels offer a means of determining the state of charge for the battery, or at least a set range of states of charge. This, in turn, helps predict battery end of life.

However, a challenge of the $Li/Ag_2V_4O_{11}$ system is the increase in internal resistance that occurs in the middle of discharge. This becomes an issue during the high current demands while charging the high voltage capacitors when it is necessary to deliver high energy shocks to the heart. As the internal resistance increases, the high power pulse voltage decreases (Fig. 9). If the voltage drop is severe enough, it can delay delivery of therapy shocks.



Medical Device Batteries. Figure 9
Low continuous current and high current pulse discharge voltages for a lithium-silver vanadium oxide ($Li/Ag_2V_4O_{11}$) cell

The cause for much of the internal resistance increase is attributed to vanadium that becomes slightly soluble in the middle portion of discharge. The dissolved vanadium diffuses to and deposits on the Li surface. If allowed to build up, the internal resistance increases [30]. The resistance can be lowered by pulling multiple high current pulses from the cell that presumably remove the vanadium as the Li dissolves during the pulse discharge [30].

Silver vanadium oxide is combined with a conductive carbon and a binder like PTFE to make the cathode. The usual electrolyte solution used is lithium hexafluoroarsenate ($LiAsF_6$) in mixed organic solvents, like propylene carbonate and 1,2-dimethoxyethane.

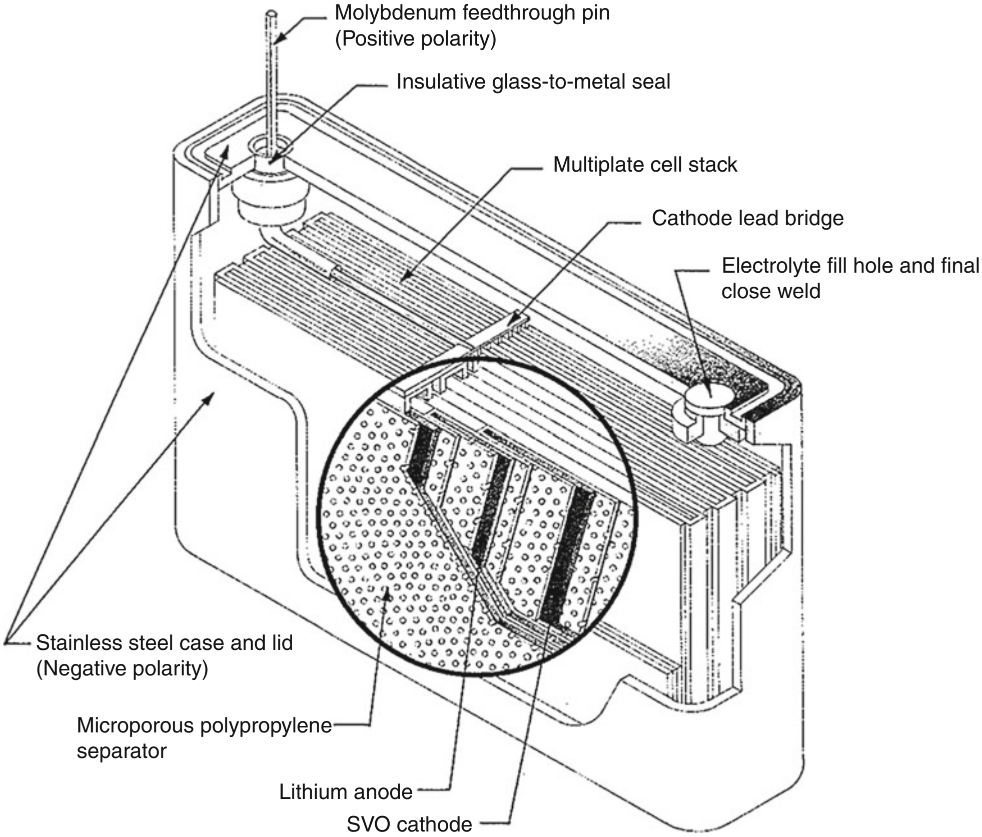
A cross section of a $Li/Ag_2V_4O_{11}$ defibrillator cell is shown in Fig. 10. Energy densities are on the order of about 270 Wh kg^{-1} and 780 Wh dm^{-3} [19].

Lithium-Manganese Dioxide The lithium-manganese dioxide (Li/MnO_2) cell was originally developed and commercialized in the mid-1970s [31] for low power applications such as wrist watches, calculators, and computer memory backup. Later, cells were developed for high power uses like cameras. Today, Li/MnO_2 cells are manufactured in a variety of sizes and configurations for low, medium, and high power consumer; military; and OEM electronic devices. Use of Li/MnO_2 cells in implantable devices was investigated for implantable cardiac pacemakers [32, 33].

Li/MnO_2 cells have a high energy density (230–270 Wh kg^{-1} and 535–520 Wh dm^{-3}) [19] that gives them the ability to provide long life for low current applications. They can also provide high power outputs required for device functions requiring high current levels. This makes Li/MnO_2 an excellent chemistry for implantable cardiac defibrillators. Indeed, Li/MnO_2 cells are now used to power implantable cardiac defibrillators from two manufacturers [34, 35].

The overall cathode reaction may be simply written as:

$$MnO_2 + xLi^+ + xe^- \rightarrow Li_xMnO_2$$

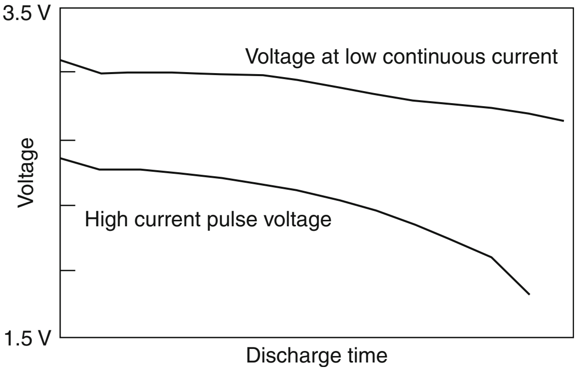

Molybdenum feedthrough pin
(Positive polarity)
Insulative glass-to-metal seal
Multiplate cell stack
Cathode lead bridge
Electrolyte fill hole and final close weld
Stainless steel case and lid
(Negative polarity)
Microporous polypropylene separator
Lithium anode
SVO cathode

Medical Device Batteries. Figure 10
Cross section of a lithium-silver vanadium oxide (Li/$Ag_2V_4O_{11}$) cell for implantable defibrillators. © Greatbatch, Inc (Reprinted with permission)

The MnO_2 positive material is similar to the type of MnO_2 used in alkaline Zn/MnO_2 cells. However, it must be heat treated at high temperatures, somewhere around 300–400°C to make it suitable for use in lithium cells. Heating removes water and modifies the structure of MnO_2. If this process is not done properly, the Li/MnO_2 cells will generate gas and internal pressure will build up, perhaps to the point where the cells will start to leak.

One or more conductive carbon powders, like acetylene black and graphite, and a binder such as PTFE are added to the MnO_2 to yield the cathode mix.

The discharge voltage of Li/MnO_2 cells exhibits two steps – rather flat during the first part of discharge, but curves downward in the last part of discharge (Fig. 11).



Medical Device Batteries. Figure 11
Low continuous current and high current pulse voltages for a Li/MnO_2 cell

Further, there is no remarkable decrease in pulse voltage as the internal resistance of the cell is relatively low and steady throughout discharge [36, 37].

Lithium-Carbon Monofluoride-Silver Vanadium Oxide One of the challenges of designing batteries for some medical devices is the wide-ranging power requirements to support different device functions. The energy density should be sufficient to provide a long life battery in a small size, yet deliver high power, particularly in ICDs and CRT-Ds.

Pairing two different positive electrode materials – one possessing a high energy density, but a low rate capability and another capable of high rate output, but a lower energy density – is a concept that has been investigated for other applications. It is now starting to be applied to medical device batteries with diverse power requirements.

A straightforward way of accomplishing this (at least from a battery design perspective) is to put together two cells with different rate capabilities, such as Li/I_2 for low rate operations and Li/MnO_2 for high rate functions [38].

Another method is to combine the two positive electrode materials in the same cell. For example, some implantable cardiac defibrillators and pacemakers have started to use lithium cells containing both CF_x (higher energy density but a lower rate capability) and $Ag_2V_4O_{11}$ (high rate capability) positive electrode materials. The CF_x discharges at a voltage that is greater than $Ag_2V_4O_{11}$ throughout most of the cell life, when the device requires only low power from the cell, which is most of the time for implantable defibrillators and pacemakers. When higher power is occasionally required from the cell, such as during RF telemetry or charging high voltage capacitors, the CF_x cannot support the higher load but the $Ag_2V_4O_{11}$ electrode material can.

Dual CF_x-$Ag_2V_4O_{11}$ positive electrodes can be made by mixing the individual CF_x and $Ag_2V_4O_{11}$ materials in the same cathode mix [39] or assembling discrete CF_x and $Ag_2V_4O_{11}$ electrode layers to form the positive electrode [40].

Neuromodulation

Neurostimulators are much like cardiac pacemakers, except they apply small electrical signals to nerve tissue. Devices are available for spinal cord stimulation to manage chronic pain, vagal nerve stimulation to control epilepsy and deep depression, and deep brain stimulation to help relieve symptoms of Parkinson's disease and other neurological disorders.

According to the US Center for Disease Control and Prevention, in 2009 more than 25% of adults, 18 years and older, have experienced low back pain for 24 h or longer [41]. The treatment for chronic pain that does not respond well to drug therapy using implantable electrical neurostimulator devices is growing.

While primary batteries have been used for these devices in the past, the trend is toward using rechargeable batteries to extend the longevity of the device. For implantable devices, the longer the battery lasts, fewer surgeries are required to replace the device, which in turn means lower risks of surgical complications, like infections.

Neuromodulation Batteries

Lithium Ion Secondary, or rechargeable, cells have been used in medical devices for some time. One of the first batteries to be used for implantable cardiac pacemakers was a custom rechargeable nickel-cadmium battery [6]. Systems to recharge the battery transcutaneously using inductive methods were developed starting in the late 1950s [42].

Some implantable medical devices use secondary cells as a way to minimize the number of device replacement surgeries compared to primary cells. An example is neurostimulation for pain management for which primary cells have almost been supplanted entirely by rechargeable cells. The result is longer device longevity in a small size.

Since they were first commercialized in 1991, lithium ion (Li ion) cells now have become the system of choice for high-end portable consumer applications (e.g., laptop and tablet computers, personal multimedia players, and cell phones). They are also seeing increased use in power tools, electric vehicles, and space satellites. Likewise, Li ion cells play an important role in some implantable medical devices.

The chemistry systems in the Li ion cells that are manufactured specifically for implantable medical devices are similar to those developed for consumer applications.

Reliability and safety are utmost concerns when designing Li ion cells for implantable medical devices. They are hermetically sealed and may contain redundant safety features in the batteries (as well as part of the charging circuitry in the device). Also, cell sizes and shapes are often specially designed for a particular device.

The negative electrode material for most Li ion cells is some form of carbon, such as graphite or coke. Cells that use lithium titanate ($Li_4Ti_5O_{12}$) negative electrode material are also available. These cells have cell voltages over 1 V lower than more traditional Li ion cells that use carbon negative electrode materials, though. The advantages are rapid charging, and they can provide thousands of charge–discharge cycles.

Unlike lithium metal, which oxidizes to soluble Li^+ during discharge of lithium primary cells, lithium ions are instead reversibly inserted, or intercalated, between graphite layers during the cell charging process and removed during discharge:

$$C_6 + xLi^+ + xe^- \underset{\text{discharge}}{\overset{\text{charge}}{\rightleftarrows}} C_6Li_x$$

The value for x is generally between 0 and 1. Carbons intercalated with Li^+ can have electrochemical potentials within tens of millivolts of the lithium metal potential when fully charged.

The structure of the carbon material remains intact after the complete charge and discharge cycle and so is able to repeat the charge–discharge process many times.

Positive electrode materials available for Li ion cells today include metal oxide or phosphate compounds, such as

- Lithium cobalt oxide ($LiCoO_2$)
- Lithium mixed metal oxide, which is a combination of cobalt (Co) and nickel (Ni) and perhaps other metal ions. Examples of mixed metal oxides are $LiNi_{0.8}Co_{0.2}O_2$, $LiNi_{0.8}Co_{0.15}Al_{0.5}O_2$, and $LiNi_{0.33}Co_{0.33}Mn_{0.33}O_2$
- Lithium manganese oxide ($LiMn_2O_4$), sometimes referred to as spinel after its structure type
- Lithium iron phosphate ($LiFePO_4$)

The first commercialized Li ion cells used $LiCoO_2$ as the positive electrode material. It consists of Li^+ inserted between the cobalt and oxygen layers. The charge and discharge reactions are similar to carbon in that Li^+ is removed and inserted leaving the basic $LiCoO_2$ unchanged:

$$LiCoO_2 \underset{\text{discharge}}{\overset{\text{charge}}{\rightleftarrows}} Li_{1-x}CO_2 + xLi^+ + xe^-$$

where x lies somewhere between 0 and 1.

The overall cell reactions are:

$$C_6 + LiCoO_2 \underset{\text{discharge}}{\overset{\text{charge}}{\rightleftarrows}} Li_xC_6 + Li_{1-x}CoO_2$$

Some positive electrode materials have three-dimensional rather than layered structures. $LiMn_2O_4$ and $LiFePO_4$ are in this category. Lithium ions are inserted in tunnels rather than between layers.

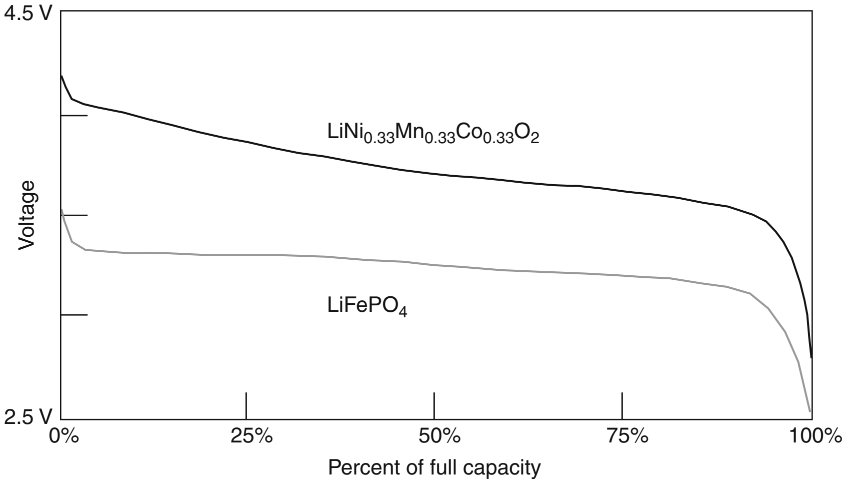
Positive electrode materials are generally mixed with a conductive carbon powder and a polymer binder. Commonly used binders are polyvinylidene fluoride (PVDF) and a PVDF copolymer with hexafluorpropylene (PVDF-HFP). The positive electrode mix is coated as a slurry onto a thin aluminum foil current collector while the carbon negative electrode material, often using the same binders, is coated onto a thin copper foil current collector.

Cylindrical or prismatic spiral wound cells are assembled by winding the positive and negative electrodes together with a porous membrane separator between them. The electrolyte solution is typically a mixture of organic solvents containing a lithium salt, like lithium hexafluorophosphate ($LiPF_6$).

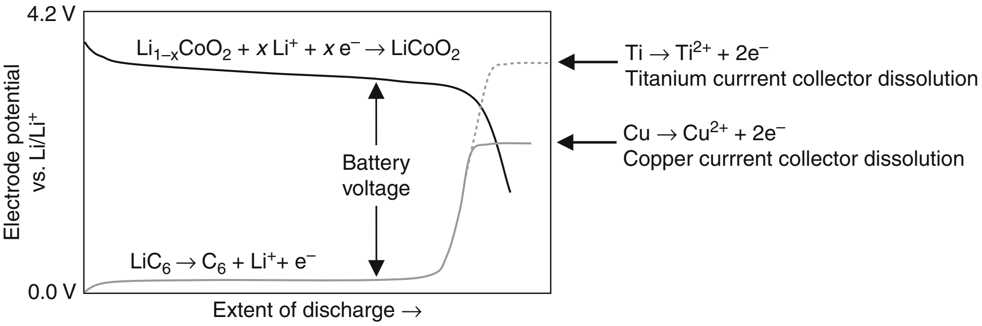
A related Li ion cell type is the Li ion polymer cell that uses a polymer gel electrolyte. Li ion polymer cells are often designed as flat prismatic cells and can be made quite thin.

The discharge voltage can be sloped or flat depending on the type of cell chemistry used (for example, see Fig. 12).

Maximum voltages at the end of charge are usually in the range of 3.6–4.2 V depending on the chemistry. Energy densities can be as high as 240 Wh kg^{-1} and 640 Wh dm^{-3}.



Medical Device Batteries. Figure 12
Discharge voltage for two types of Li ion cells



Medical Device Batteries. Figure 13
Individual electrode potentials for a lithium ion cell

Developers of secondary cells for implantable medical devices must consider the possibility of what happens when the battery is not charged for whatever reason in a timely fashion. If the cell becomes over discharged and its voltage is allowed to go below a certain threshold, the cell may not fully recover once it is charged and performance may be reduced.

One reason cell performance may be compromised in the event of the issues with typical Li ion cells is the use of a copper (Cu) current collector for the negative electrode. As a cell is discharged, the positive electrode potential decreases and the negative electrode potential increases. When the cell voltage drops to about 0 V, the potential of the negative electrode is the same as the positive electrode when measured against a reference electrode like Li. If discharge continues, the negative electrode potential could reach the potential at which the Cu current collector oxidizes or corrodes and copper ions dissolve (Fig. 13) [43]. This can lead to loss of cell capacity and degrade cell function. For this reason, the cell anode to cathode capacity balance is a critical design consideration. Cell performance is limited by the cathode capacity.

Selection of a positive electrode material that discharges at a lower voltage can minimize this issue, but an alternative is to replace the Cu current collector with a metal that corrodes at a more positive potential, such as titanium (Ti) [43]. In this case, the capacity of a cell that is held at 0 V for prolonged periods of time can retain more of its capacity [44].

Hearing Loss

The US National Institutes of Health (NIH) reports that 36 million adults in the US report some level of hearing loss [45]. The World Health Organization puts the number at 278 million worldwide in 2005 [46]. Up to 0.3% of children born in the USA are deaf or hard of hearing [45].

Hearing aids are small devices that amplify the sounds picked up by a tiny microphone to enable the hearing-impaired to function. There are several types of hearing aid designs: behind the ear, in the ear, and completely in the canal. NIH finds that 1 out of 5 people in the USA who could benefit from a hearing aid actually wears one [45]. Worldwide, it is less than 1 out of 40 people [46].

The first portable radios operated by batteries were developed in the 1930s. The invention of the first portable hearing aids followed soon after in 1937 [47]. It could be carried by the hearing aid user, but since it used vacuum tubes it was the size of a lunch box. Wearable hearing aids became possible with the development of smaller electronic components that replaced vacuum tubes and the commercialization of small zinc-mercury oxide cells after World War II.

Most hearing aids today run on zinc-air cells because they have a high energy density and so last longer. Depending on the hearing needs of the user and the type of hearing aid, batteries may last from a few days to over 1 month. Recently, rechargeable nickel-metal hydride and zinc-silver oxide cells hearing aid batteries have been introduced.

Hearing Aid Batteries

Zinc-Air The design of a zinc-air cell is different than most other battery types. Nearly all cells store all of the active materials required for the cell to function – negative and positive electrode materials, electrolyte – within a battery case or housing.

However, the positive electrode material for zinc-air cells is atmospheric oxygen, O_2. A thin electrode, or cathode, similar to a fuel cell electrode that provides a site for O_2 to be reduced during cell discharge is all that is necessary. As a result, there is more space within the cell that can be loaded with more of the zinc (Zn) negative electrode material.

The combination of a practically unlimited source of O_2 from the air with more available space inside the cell for additional Zn results in a very high energy density cell. Zinc-air button cells have the highest energy density of all commonly available button cells – up to 370 Wh kg^{-1} and 1,300 Wh dm^{-3} [14]. Of course, using atmospheric O_2 means zinc-air cells cannot be used for implantable devices.

Air was known to improve cell performance early in the nineteenth century, even though the electrochemical mechanism was unknown. Cells that used atmospheric O_2 as the positive electrode material were built in the late 1800s. Practical advances in 1930s through the 1950s made zinc-air cells commercially viable. Prismatic zinc-air cells were developed to operate hearing aids in the early 1950s [48, 49]. The early hearing aid cells were close to 22 cm^3 in volume with energy densities of about 200 Wh kg^{-1} [50].

Improvements in hearing aid electronics resulted in smaller devices that required less power. Lower power demands and advances in battery technology lead to smaller batteries, as well. Typical primary hearing aid cells today are small button cells that range from 5.8 mm in diameter and 2.15 mm in height to 11.5 mm in diameter and 5.4 mm in height (Table 1).

The largest hearing aid cell today is less than about 0.6 cm^3 – almost forty times less volume than the first zinc-air hearing aid batteries.

The zinc is in powder form that is mixed with a compound like polyacrylic acid or sodium carboxymethyl cellulose that is chemically stable and forms a gel in the alkaline electrolyte solution. The gelled anode is more stable mechanically.

The discharge reaction of Zn in aqueous potassium hydroxide (KOH) electrolyte solutions is somewhat complex, but may be simply written as:

$$Zn + 4OH^- \rightarrow Zn(OH)_4{}^{2-} + 2e^-$$

where the zincate product, $Zn(OH)_4{}^{2-}$, is dissolved in the electrolyte. Under certain conditions, especially as the battery is depleted, zinc ions can precipitate as zinc oxide (ZnO):

$$Zn + 2OH^- \rightarrow ZnO + H_2O + 2e^-$$

Typical alkaline, or basic, electrolyte solutions are comprised of 20–50% KOH (typically 28%) by weight

Medical Device Batteries. Table 1 Common primary hearing aid button cell sizes. IEC is the International Electrotechnical Commission and ANSI is the American National Standards Institute

Size designation			Dimensions		Available cells types (IEC, ANSI)
IEC	ANSI	Common	Diameter (mm)	Height (mm)	
63	7012ZD	5	5.8	2.15	Zinc-air (PR63, 7012ZD)
70	7005ZD	10	5.8	3.6	Zinc-air (PR63, 7012ZD)
	1191SO				Zinc-silver oxide (SR70, 1191SO)
41	7002ZD	312	7.9	3.6	Zinc-air (PR41, 7002ZD)
	1135SO				Zinc-silver oxide (SR41, 1135SO)
48	7000ZD	13	7.9	5.4	Zinc-air (PR48, 7000ZD)
	1137SO				Zinc-silver oxide (SR48, 1137SO)
44	7003ZD	675	11.6	5.4	Zinc-air (PR44, 7003ZD)
	1131SO				Zinc-silver oxide (SR44, 1131SO)

in water. A small amount of ZnO, between a few percent up to saturation levels, is added to the KOH electrolyte to help reduce Zn self-discharge. ZnO dissolves to form potassium zincate in the electrolyte.

Zinc is used as the negative electrode material in aqueous primary cells because it has a high energy density. However, Zn is prone to corrosion, particularly when certain impurities are present, that can result in decreased battery life and compromised performance. A common contaminant is iron. When it comes into contact with the Zn and becomes galvanically coupled with it, hydrogen gas (H_2) will form at the iron surface from the reduction of water in the electrolyte. Zinc is oxidized and dissolves into the electrolyte as zincate or precipitates as ZnO.

If Zn corrosion proceeds at a sufficiently high rate for long periods of time, the gas pressure within the cell can build until the seals start to leak gas and electrolyte. This was a common problem in early cardiac pacemakers using Zn/HgO batteries. The gassing cannot be entirely avoided, but the corrosion reactions can be slowed and whatever gas is formed can be controlled to maximize the shelf life and minimize the loss of cell performance.

Alloying Zn with mercury metal to form an amalgam increases the electrochemical overpotential for hydrogen gas formation on Zn. Zinc amalgams in the range of 2–15% Hg by weight of Zn were common.

Since mercury (Hg) and its compounds are toxic, beginning in the early 1990s the European Union, followed by the United States, required battery manufacturers to reduce [51] and then eliminate [52] Hg added to batteries. This requirement addressed genuine concerns that cells discarded in landfills could release Hg into the environment.

Use of battery materials with higher purity levels (including Zn, electrolyte, and cathode), different alloying elements, seals that can withstand higher internal pressures, and cleaner manufacturing processes enabled the elimination of added Hg in batteries.

The positive electrode reaction involves the reduction of O_2 and proceeds in two basic steps:

$$O_2 + H_2O + 2e^- \rightarrow HO_2^- + OH^-$$

and then

$$2HO_2^- \rightarrow 2OH^- + O_2.$$

where HO_2^- is the hydrogen peroxide ion. The second step acts to curb the oxygen discharge reaction, so a catalyst, usually manganese dioxide, may be used to accelerate peroxide decomposition.

The overall cell reaction is:

$$2Zn + O_2 + 4KOH + 2H_2O \rightarrow 2K_2Zn(OH)_4$$

Or

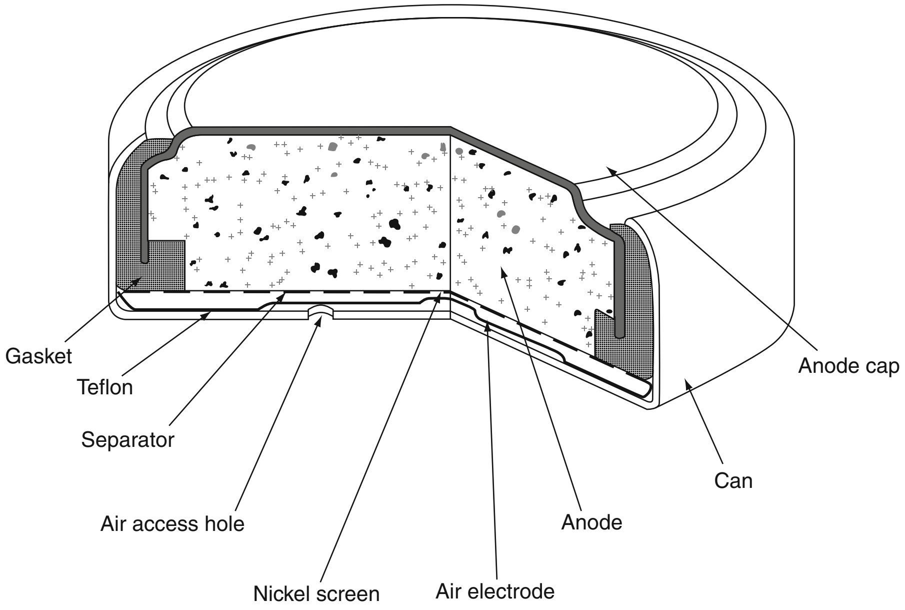
$$2Zn + O_2 \rightarrow 2ZnO$$

The reduction of O_2 is thermodynamically favorable, but the reaction kinetics are relatively slow. Using an electrode material with a high surface area, such as activated carbon powder, can overcome the slow oxygen reduction kinetics by presenting a large number of sites at which O_2 can be reduced. A small amount of a conductive carbon material may also be added to increase the electrical conductivity of the cathode.

While it is important to maximize the electrode area accessible to O_2, the cathode must also allow for contact with the electrolyte. A three-phase boundary comprised of O_2, electrolyte and carbon with catalyst is central to the proper functioning of an air electrode. Maintaining a balance between access to atmospheric O_2 and electrolyte solution is achieved by adequately dispersing just the right amount of a hydrophobic material, polytetrafluoroethlyene (PTFE or Teflon®), in the carbon powder mix. Too little PTFE or it is not dispersed well enough and the electrolyte solution could saturate the cathode. This reduces the area at which O_2 can react with a resulting loss of performance. Too much PTFE and the electrolyte will not sufficiently wet the cathode. Again, this limits battery performance.

A zinc-air cathode must provide a surface at which O_2 can be reduced during battery discharge while not being consumed itself during cell discharge. The zinc-air cathode of today is comprised of multiple layers, each serving an important function. A carbon mix (activated carbon, conductive carbon, and PTFE) is applied to a nickel screen current collector that carries electrons to the cathode during cell discharge. There are two sides to an air cathode – the air side and the electrolyte side. The air side is laminated with a porous PTFE membrane. A polymeric separator sheet is pressed on the electrolyte.

Important considerations for zinc-air cells are air and water management. Air must be allowed to enter the zinc-air cell, which is done by providing holes in the battery can near the cathode (see Fig. 14). There must be an adequate number of holes of a sufficient size to allow enough air to enter the cell such that the discharge reaction is not impeded by lack of O_2. An air diffusion layer may be included just inside the cell to disperse the air more uniformly.

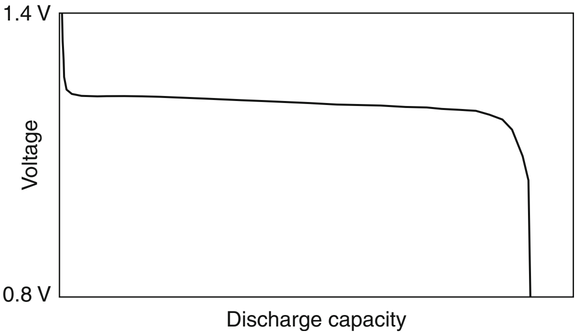


Medical Device Batteries. Figure 14
Cross section illustrating the main components of a zinc-air cathode. © Eveready Battery Company, Inc (Reprinted with permission)

The advantage of zinc-air cells also presents a big challenge for battery designers. Remaining open to the atmosphere renders zinc-air cells exposed to detrimental environmental conditions, especially humidity. Water in humid air can be absorbed by the basic electrolyte solution diluting it and subsequently flooding the cathode. Arid air may evaporate water from the electrolyte and dry the cathode. Both conditions lead to reduced cell performance and battery life. Carbon dioxide in the air can enter the cell, react with the basic electrolyte solution and precipitate carbonates, also decreasing performance.

A simple way to mitigate the effects of environmental exposure is to seal the holes until the cell is needed to operate. The holes are covered, usually with an adhesive tab, after manufacture to curtail exposure to the atmosphere during shipping and storage. The seal is removed prior to use, air enters the cell and it is ready for use. Hearing aid users can also replace the adhesive tab between uses, overnight while sleeping for example, to extend cell life.

The initial open circuit potential of the zinc-air cell is about 1.45 V. The voltage of a zinc-air cell is mostly constant throughout discharge (Fig. 15).



Medical Device Batteries. Figure 15
Discharge voltage behavior for a zinc-air cell

Zinc-Silver Oxide Zinc-silver oxide (Zn/Ag_2O) cells are used in medical applications like hearing aids, but also military, aerospace, watches, cameras, and calculators. Early implanted electrical bone growth stimulators [53] for fracture healing and spinal fusion used Zn/Ag_2O batteries. The Zn/Ag_2O cell is one of the alternatives to Zn/HgO cells (see above) and is used in applications where high energy and power density is required.

Originally dating back to 1883, [54] alkaline Zn/Ag_2O cells were eventually developed into commercially practical cells in the early 1960s.

The negative electrode material is zinc powder in a gelled KOH-ZnO-H_2O alkaline electrolyte solution. NaOH can be used instead of KOH for lower power applications.

The positive electrode reaction is:

$$Ag_2O + H_2O + 2e^- \rightarrow 2Ag + 2OH^-$$

The cathode pellet contains Ag_2O powder and 1–5% of a conductive carbon powder like graphite, to reduce internal resistance and provide good contact to all of the active silver oxide particles, mixed with a PTFE binder to maintain the mechanical integrity of the pellet.

A disadvantage of Ag_2O is its solubility in alkaline electrolyte. Silver ions dissolve into basic electrolyte solutions. The solubility of silver ions (Ag^+) from Ag_2O is on the order of 10^{-4} mol dm^{-3} in concentrated KOH [55]. Dissolved Ag^+ diffuses to the Zn and deposits there as Ag. If this continues, the Ag deposit will grow as dendrites through the separator and eventually create an internal short circuit by directly bridging the positive and negative electrode materials [56]. A big advance toward making this cell chemistry practical came when Henri André developed a cellophane separator that minimized diffusion of Ag^+ through the separator which mitigated this issue [54].

The overall cell reaction is:

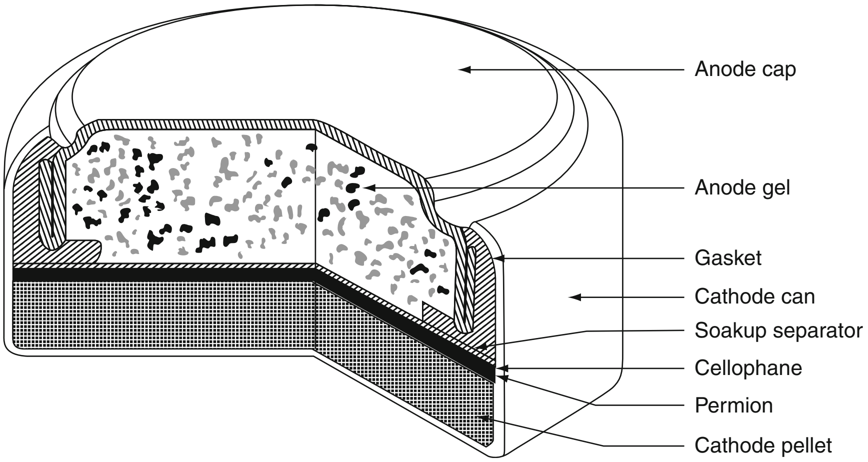
$$Zn + Ag_2O + 2KOH + H_2O \rightarrow K_2Zn(OH)_4 + 2Ag$$

or

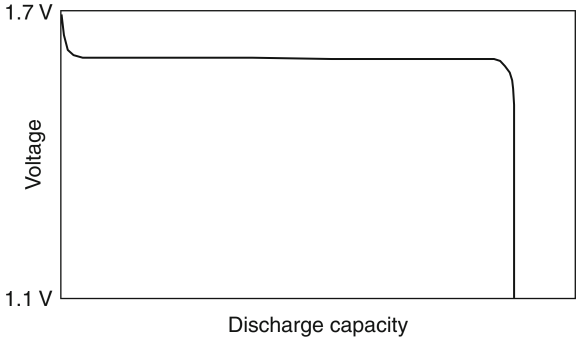
$$Zn + Ag_2O \rightarrow ZnO + 2Ag$$

The cross section of a Zn/Ag_2O button cell is pictured in Fig. 16.

The open circuit potential of Zn/Ag_2O cells is about 1.60 V and voltage remains relatively constant throughout discharge as seen in Fig. 17. Energy density for Zn/Ag_2O button cells is about 135 Wh kg^{-1} or 530 Wh dm^{-3} [14].



Medical Device Batteries. Figure 16
Cross section illustrating the main components of a Zn/Ag_2O button cell. © Eveready Battery Company, Inc (Reprinted with permission)



Medical Device Batteries. Figure 17
Discharge voltage behavior for a Zn/Ag_2O cell

Future Directions

A wide variety of wearable, implantable, and even ingestible medical devices are under development or in the early stages of clinical use. Here are just a few.

There is an ongoing need to treat various heart conditions, hearing loss, and chronic pain. Improvements to electronic circuit components and designs, along with the batteries that make them function, will continue. New therapies and device features will be enabled by new battery technologies.

Additionally, there are many emerging indications for wearable or implantable medical devices, particularly new neuromodulation applications such as deep brain stimulation for various movement and neurological disorders [57] and occipital nerve stimulation to treat migraine and cluster headaches [58]. Implantable visual prostheses to restore sight are also under development.

There are ingestible devices for measurement of body core temperature and other sensors to help diagnose gastrointestinal tract disorders. Implantable sensors [59, 60], such as those for blood pH, oxygen, and glucose, along with remote telemetry, are expected to advance as well.

Bibliography

Primary Literature

1. Althaus J (1860) A treatise on medical electricity, theoretical and practical. Lindsay and Blakiston, Philadelphia, pp v–vi
2. Tibbits H (1873) A handbook of medical electricity. P. Blakiston, Son and Co., Philadelphia, p v
3. Steavenson WE, Jones HL (1892) Medical electricity. A practical handbook for students and practioners. P. Blakiston, Son and Co., Philadelphia, pp v–vi
4. Mohrman DE, Heller LJ (1996) Cardiovascular physiology, 4th edn. McGraw-Hill, New York
5. Lillehei CW, Gott VL, Hodges PC, Long DM, Bakken EE (1960) Transistor pacemaker for treatment of complete atrioventricular dissociation. JAMA 172:2006–2010
6. Greatbatch W, Holmes CF (1991) History of implantable devices. IEEE Eng Med Biol:38–41,49

7. Chardack W, Gage A, Greatbatch W (1960) A transistorized, self-contained, implantable pacemaker for the long-term correction of complete heart block. Surgery 48:543
8. Furman S (2003) The early history of cardiac pacing. Pacing Clin Electrophysiol 26:2023–2032
9. Liang CC, Holmes CF (1980) Lithium pacemaker batteries – an overview. In: Owens BB, Margalit N (eds) Proceedings of the symposia on power sources for biomedical implantable applications and ambient temperature batteries, vol 80–84, pp 27–33
10. Bilitch M, Parsonnet V, Furman S (1980) Clinical assessment of pacemaker power sources. In: Owens BB, Margalit N (eds) Proceedings of the symposia on power sources for biomedical implantable applications and ambient temperature batteries, vol 80–84, pp 18–26
11. Greatbatch W (1984) Pacemaker power sources. IEEE Eng Med Biol Mag:15–19
12. Greatbatch W, Lee JH, Mathias W, Eldridge M, Moser JR, Schneider AA (1971) The solid-state lithium battery: a new improved chemical power source for implantable cardiac pacemakers. IEEE Trans BioMed Eng BME 18:317–324
13. Kelly RG, Moran PJ (1987) The rate limiting mechanism of Li/I_2 (PV2P) batteries. J Electrochem Soc 134:25–30
14. Linden D (2002) Ch 7 Primary batteries – introduction. In: Linden D, Reddy TB (eds) Handbook of batteries, 3rd edn. McGraw-Hill, New York
15. Shellock FG, Hatfield M, Simon BJ, Block S, Wamboldt J, Starewicz PM, Punchard WFB (2000) Implantable spinal fusion stimulator: assessment of MR safety and artifacts. J Mag Res Imaging 12:214–223
16. Maurer J, Savvas E (2010) The Esteem system: a totally implantable hearing device. In: Böheim K (ed) Active middle ear implants. S Karger AG, Basel, p 59
17. Root MJ, Dumas R, Yazami R, Hamwi A (2001) The effect of carbon starting material on carbon fluoride synthesized at room temperature – characterization and electrochemistry. J Electrochem Soc 148:A339–A345
18. Spellman PJ, Dittberner KL, Pilarzyk JG, Root MJ (2000) Application of Li/CF_x system in portable electronics. In: Osaka T, Datta M (eds) Energy storage systems for electronics. Gordon and Breach Science, Amsterdam, pp 131–152
19. Linden D, Reddy TB (2002) Ch 14 Lithium batteries. In: Linden D, Reddy TB (eds) Handbook of batteries, 3rd edn. McGraw-Hill, New York
20. Parker B (1978) Obituary: a vindication of the zinc-mercury pacemaker battery. Pacing Clin Electrophysiol 1:148–149
21. Clarke CL (1884) US patent 298,175
22. Brønsted JN (1917) US patent 1,219,074
23. Friedman M, McCauley CE (1947) The Ruben cell – a new alkaline primary dry cell battery. Trans Electrochem Soc 92:195–215
24. Ruben S (1978) The evolution of electric batteries in response to industrial needs. Dorrance, Philadelphia, Ch VI
25. Mirowski M, Reid PR, Mower MM, Watkins L, Gott VL, Schauble JF, Langer A, Heilman MS, Kolenik SA, Fischell RE, Weisfeldt ML (1980) Termination of malignant ventricular arrhytmias with an implanted automatic defibrillator in human beings. N Engl J Med 303:322–324
26. Liang CC, Bolster ME, Murphy RM (1982) US patent 4,310,609
27. Leising RA, Takeuchi ES (1993) Solid-state cathode materials for lithium batteries:effect of synthesis temperature on the physical and electrochemical properties of silver vanadium oxide. Chem Mater 5:738–742
28. Crespi A (1993) US patent 5,221,453
29. Leising RA, Thiebolt WC, Takeuchi ES (1994) Solid state characterization of reduced silver vanadium oxide from the Li/SVO discharge reaction. Inorg Chem 33:5733–5740
30. Syracuse K, Waite N, Gan H, Takeuchi ES (2006) US patent 6,982,543
31. Ikeda H, Saito T, Tamura H (1975) Manganese dioxide as cathodes for lithium batteries. In: Proceedings of the manganese dioxide symposium, vol 1. The Electrochemical Society, Cleveland, pp 384–401
32. Gerbier G, Lehman G (1980) Mangalith: a new lithium pacemaker battery. In: Owens BB, Margalit N (eds) Proceedings of the symposia on power sources for biomedical implantable applications and ambient temperature batteries, vol 80–84. The Electrochemical Society, Pennington, pp 136–143
33. Merritt DR, Schmidt CL (1993) Impedance modeling of the lithium/manganese dioxide battery. In: Surampudi S, Koch VR (eds) Proceedings of the symposium on lithium batteries, vol 93–24. The Electrochemical Society, Pennington, pp 138–145
34. Drews J, Fehrmann G, Staub R, Wolf R (2001) Primary batteries for implantable pacemakers and defibrillators. J Power Sources 97–98:747–749
35. Root MJ (2008) Implantable cardiac rhythm device batteries. J Cardiovasc Trans Res 1:254–257
36. O'Phelan MJ, Victor TG, Haasl BJ, Swanson LD, Kavanagh RJ, Barr AG, Dillon RM (2009) US patent 7,479,349
37. Root MJ (2010) Lithium-manganese dioxide cells for implantable defibrillators – discharge voltage models. J Power Sources 195:5089–5093
38. Drews J, Wolf R, Fehrmann G, Straub R (1999) Development of a hybrid battery system for an implantable biomedical device, especially a defibrillator/cardioverter (ICD). J Power Sources 80:107–111
39. Weiss DJ, Cretzmeyer JW, Crespi AM, Howard WG, Skarstad PM (1993) US patent 5,180,642
40. Spillman DM, Takeuchi ES (1999) US patent 5,935,724
41. National Center for Health Statistics (2010) Health, United States, 2009. US Department of Health and Human Services Center for Disease Control and Prevention, Publication 2010–1232, Hyattsville, p 261
42. Fischell RE, Schulman JH (1976) A rechargeable power system for cardiac pacemakers. Proc 11th IECEC, pp 163–168
43. Tsukamoto H, Kishiyama C, Nagata M, Nakahara H, Piao T (2003) US patent 6,596,439
44. Kishiyama C, Nagata M, Piao T, Dodd J, Lam P, Tsukamoto H (2003) 204th Electrochemical Society Meeting, abstract 425

45. http://wwwnidcdnihgov/health/statistics/quickhtm. Accessed 27 Dec 2010
46. http://wwwwhoint/mediacentre/factsheets/fs300/en/indexhtml. Accessed 3 Jan 2011
47. Wengel AM (1940) US patent 2,192,669
48. Marsal PA, Fox RP (1952) US patent 2,597,116
49. Schumacher EA, Bennett RJ (1952) US patent 2,597,119
50. Brodd RJ, Kozawa A, Kordesch KV (1978) Primary batteries 1951–1976. J Electrochem Soc 125:271C–282C
51. Council Directive 91/157/EEC of 18 March 1991 on Batteries and Accumulators Containing Certain Dangerous Substances
52. "NEMA announces battery industry commitment to eliminating mercury in button cells". https://wwwnemaorg/media/pr/20060302acfm. Accessed 28 Dec 2010
53. Paterson DC, Lewis GN, Cass CA (1980) Treatment of delayed union and non-union with an implanted direct current stimulator. Clin Orthop Relat Res 148:117–128
54. Howard PL, Fleischer A (1971) Ch 1. Milestones in the electrochemistry of zinc-silver oxide batteries. In: Fleischer A, Lander JJ (eds) Zinc-silver oxide batteries. Wiley, New York
55. Amlie RF, Reutschi P (1961) J Electrochem Soc 108:813–819
56. Himy A (1986) Silver-zinc battery phenomena and design principles. Vantage, New York, p 7
57. Pereira EAC, Aziz TZ (2006) Surgical insights into Parkinson's disease. J R Soc Med 99:238–244
58. Schwedt TJ, Dodick DW, Trentman TL, Zimmerman RS (2006) Occipital nerve stimulation for chronic cluster headache and hemicrania continua: pain relief and persistence of autonomic features. Cephalagia 26:1025–1027
59. Van Laerhoven K, Lo BPL, Ng JWP, Thiemjarus S, King R, Kwan S, Gellersen H-W, Sloman M, Wells OW, Needham P, Peters N, Darzi A, Toumazou C, Yang G-Z (2004) Medical healthcare monitoring with wearable and implantable sensors. Proc 3rd international workshop on ubiquitous computing for healthcare applications. http://www.pervasivehealthcare.com/ubicomp2004/papers/final_papers/laerhoven.pdf
60. Fletter PC, Majerus S, Cong P, Damaser MS, Ko WH, Young DJ, Garverick SL (2009) Wireless micromanometer system for chronic bladder pressure monitoring. 6th international conference on networked sensing systems, IEEE, pp 1–4

Books and Reviews

Gabano J-P (ed) (1983) Lithium batteries. Academic, London

Holmes CF (1994) Implantable lithium power sources. In: Pistoia G (ed) Lithium batteries. Elsevier, Amsterdam, pp 377–416

Owens BB (ed) (1986) Batteries for implantable biomedical devices. Plenum, New York

Reddy TB (ed) (2010) Linden's handbook of batteries, 4th edn. McGraw-Hill, New York

Root M (2010) The TAB battery book. McGraw-Hill, New York

Schlesinger H (2010) The battery. HarperCollins, New York

Takeuchi ES, Leising RA, Spillman DM, Rubino R, Gan H, Takeuchi KJ, Marschilok AC (2004) Lithium batteries for medical applications. In: Nazri G-A, Pistoia G (eds) Lithium batteries science and technology. Kluwer, Boston, pp 686–700

Untereker DF, Crespi AM, Rorvick A, Schmidt CL, Skarstad PM (2007) Power systems for implantable pacemakers, cardioverters, and defibrillators. In: Ellenbogen KA, Kay GN, Lau C-P, Wilkoff BL (eds) Clinical cardiac pacing and defibrillation, 3rd edn. Saunders, Philadelphia, pp 235–259

Vincent C, Scrosati B (1997) Modern batteries, 2nd edn. Arnold, London

Medicinal Plants, Engineering of Secondary Metabolites in Cell Cultures

Suvi T. Häkkinen, Anneli Ritala, Heiko Rischer, K.-M. Oksman-Caldentey
VTT Technical Research Centre of Finland, Espoo, Finland

Article Outline

Glossary

Bioreactor A fermentor in which plant cell cultures can be cultivated in sterile, controlled, and contained condition for biotechnological production of cell biomass and/or particular protein or small molecule.

Medicinal plants Plants that are used for medicinal purposes; whole plants or specific plant organs or compounds derived thereof can be utilized.

Metabolic engineering A process to understand metabolic pathways; a targeted alteration of metabolic pathways with the aim of improved yield, quality, and/or spectrum of produced metabolites.

Plant cell culture Process where plant cells are cultivated under controlled conditions; may consist of differentiated tissues or organs (e.g., shoots, roots, embryos, stems) or undifferentiated cells (e.g., callus, suspension cultures).

Secondary metabolites Low molecular weight compounds with enormous chemical diversity often found in plants in small amounts essential for plants' defense system; many secondary metabolites are used as pharmaceuticals, dyes, flavors, and fragrances by humans.

Transgene A gene that has been transferred from one organism to another.

Definition of the Subject

Plants are the most excellent designers and producers of a variety of small compounds that are beneficial to mankind as foods, medicines, and industrial raw materials. The use of medicinal plants for human health dates back to ancient history of mankind. The first written document of the use of medicinal plants can be found in Papyrus Ebers (1800 BC). Even if the use of certain medicinal plants was known to treat certain diseases – often using the trial-and-error approach – it is only less than 200 years ago the isolation of the first active chemical constituent (secondary metabolite) responsible for its pharmacological effect occurred. Today, many plant-derived compounds are used in pharmaceutical industry, and plants also serve as an important source for new lead compounds.

Many plants containing high-value secondary metabolites are difficult to cultivate or are becoming endangered because of the overharvesting. Furthermore, the chemical synthesis of plant-derived compounds is often not economically feasible due to their highly complex structures and the specific stereochemical requirements of the compounds. The biotechnological production of valuable secondary metabolites in plant cell or organ cultures is an attractive alternative to the extraction of whole plant material. However, the use of plant cell or organ cultures has had only limited commercial success so far. This is explained by the empirical nature of selecting high-yielding, stable cultures and the lack of understanding of how secondary metabolites are synthesized or how their synthesis is regulated.

Introduction

It has been estimated that there are at least 400,000 higher plant species in the world of which only about 10% are characterized chemically to certain extent [1]. There is no doubt that the chemical diversity of plants is much greater than any chemical library made by humans, and thus the plant kingdom represents an enormous reservoir of pharmacologically valuable molecules waiting to be discovered. Plants are thus excellent organic chemists in nature and constantly respond to environmental changes by adjusting their capacity to produce natural products. Functional genomics may open entirely new avenues to screen unexplored medicinal plant species for their pharmacological value. Many pharmaceutical companies have now renewed their interest on plant-derived compounds due to too high expectations on combinatorial chemistry or computational drug design to obtain new drug leads during the past decades [2, 3].

Many secondary metabolites of industrial value are complex in their structures making chemical synthesis very challenging and expensive. Moreover, plants contain usually very low contents of these compounds, and therefore other production processes are essential to be developed. Biotechnological production using plant cells as real green factories is a very promising technology, but currently there are still many limiting factors, mainly related to our poor understanding how the plants synthesize these high-value compounds and how the synthesis is regulated.

In the following sections, an overview is given how secondary metabolites are produced in plant and tissue cultures, how the production can be enhanced by classical optimization methods, and what metabolic engineering has to offer today and in the future. Spectacular advances in plant genomics and metabolite profiling offer unprecedented possibilities to explore the extraordinary complexity of the plant biochemical capacity. State-of-the-art genomics tools can be used to engineer the enhanced production of known target metabolites or to synthesize entire novel compounds by the so-called combinatorial biochemistry in cultivated plant cells. Finally, some future perspectives are given for novel techniques and tools that are just now emerging.

High-Value Products from Medicinal Plants

Medicinal Plants

Many plants such as crops play a central role in our everyday diet. The nutritional value of edible plants and their constituents has been studied for decades. Besides the edible plants, there is a huge variety of toxic plants in the plant kingdom. These include, for example, many alkaloid or terpene containing medicinal plants such as *Atropa belladonna*, *Camptotheca acuminata*, *Capsicum annuum*, *Catharanthus roseus*, *Erythroxylum coca*, *Papaver somniferum*, *Cannabis sativa*, *Artemisia annua*, and *Taxus* species – just to name a couple of them. These plants have been and still are an important source of pharmaceuticals. Molecules derived from medicinal plants make up a sizable proportion of known drugs currently available on the market. These include compounds such as morphine, codeine, and several anticancer drugs such as paclitaxel, vincristine, and vinblastine, the monetary value of which is very high. In Western medicine, over 25% of prescription drugs sold in pharmacies contain at least one active principle which is directly or indirectly (via semi-synthesis) a natural product. This number does not include the over-the-counter sold drugs or pure phytopharmaceuticals.

According to WHO, 11% of the current 252 drugs considered essential for humans are exclusively derived from flowering plants. Furthermore, plants are also important source of new drug lead compounds. During the past 25 years, 1,010 new drug entities (NDEs) were introduced to the market; 27% of them were either natural products or derived from natural products as semi-synthetic derivatives [3]. In addition, 15% of the drugs were synthesized after the molecule was first discovered from natural resources. Table 1 shows the origin of the 458 NDEs representing the four major therapy groups with anti-infectives (antibacterial, antiviral, antifungal, and antiparasitic), anticancer, antihypertensive, or anti-inflammatory activities discovered between 1981 and 2006. It is remarkable that over 68% of all antibacterial compounds and 51% of all anticancer drugs were directly or indirectly derived from natural resources. Natural sources will undoubtedly continue to play a prominent role in the discovery of pharmaceuticals in the future.

Secondary Metabolism in Plants

Secondary metabolites are low molecular weight compounds found in small quantities throughout the whole plant kingdom. They exhibit many biological functions vital for the survival of the plant such as responding to stress, mediating pollination, or acting as defense compounds. In plant cell, they are accumulated often in the vacuoles. Besides the importance for the plant itself, secondary metabolites have always been of interest to humans as flavors, fragrances, dyes, pesticides, and pharmaceuticals. However, for most of the secondary metabolites, the exact function in plants still remains unknown.

Medicinal Plants, Engineering of Secondary Metabolites in Cell Cultures. Table 1 Number of new drug entities (NDEs) discovered during 1981–2006 belonging to the four most important therapy groups (modified from [3])

Therapy group	Total	N	ND	NS	B	S	V	N+D+NS	%
Antimicrobial	230	12	74	34	13	60	37	120	52.2
Anti-bacterial	109	10	64	1	0	23	11	75	68.8
Anti-fungal	29	0	3	0	1	25	0	3	10.3
Anti-viral	78	0	2	31	12	8	25	33	42.3
Anti-parasitic	14	2	5	2	0	4	1	9	64.3
Anti-cancer	100	9	25	17	17	30	2	51	51.0
Anti-hypertensive	77	0	2	34	0	41	0	36	46.8
Anti-inflammatory	51	0	13	0	1	37	0	13	25.5
Total	458	21	114	85	31	168	39	220	48.0

N natural product, *ND* natural product derivative, *NS* product is synthesized but the original molecule is discovered from natural sources, *B* biotechnologically produced compound (often a large molecule, protein), *S* synthetic molecule, *V* vaccine

More than 200,000 secondary metabolites have hitherto been discovered from the plant kingdom, but only half of them are structurally fully elucidated [4–6]. They are characterized by an enormous chemical diversity, and every plant has its own characteristic set of secondary metabolites. The production of specific alkaloids is often strongly restricted to certain plant families, whereas, for example, flavonoids are abundant in many plant species. Based on their biosynthetic origins, plant secondary metabolites can be structurally divided into five major groups: polyketides, isoprenoids (e.g., terpenoids), alkaloids, phenylpropanoids, and flavonoids [7]. The polyketides are produced via the acetate-mevalonate pathway; the isoprenoids (terpenoids and steroids) are derived from the five-carbon precursor, isopentenyl diphosphate (IPP) produced via the classical mevalonate pathway, or the novel MEP pathway (see the details in section "Targeting the Metabolic Enzymes"); the alkaloids are synthesized from various amino acids; phenylpropanoids are derived from aromatic amino acids phenylalanine or tyrosine; and the flavonoids are synthesized by the combination of phenylpropanoids and polyketides [8].

Since the discovery of the opium alkaloid morphine almost two centuries ago, alkaloids are still one of the most studied groups of plant secondary metabolites although terpenoids are the largest chemical family of secondary metabolites. It is somehow surprising that such an extensive array of different nitrogen-containing organic molecules are known in higher plants even though only 2% of the plant dry weight is composed of the element nitrogen. The largest requirement of nitrogen is the synthesis of amino acids which function as building blocks of proteins as well as precursors to many secondary metabolites. Alkaloids are thus the most prominent nitrogenous compounds with diverse, complex structures and often possessing strong physiological properties leading their wide use as pharmaceuticals. Human use of them dates back to more than 3,000 years. Currently, more than 12,000 alkaloids are known and they are classified into several subclasses based on the amino acids from which they are derived and according to their chemical structures [9].

At the present time, small amounts of plant compounds including alkaloids, for example, morphine, scopolamine, and vincristine are isolated with often some difficulties from natural vegetation or cultivated plants which explain the high price of the raw material. Numerous secondary metabolites have also served as models for modern synthetic pharmaceuticals [3]. However, the biosynthetic pathways leading to their formation in plants are often long, complex multistep events catalyzed by various enzymes, and are still largely unknown in enzymatic and genetic level. The best characterized pathways after the decades' intensive classical biochemical research are the biosynthesis of opium and terpenoid indole alkaloids.

Besides the low quantities of the compounds in plants, the production rates may vary from year to year and secondary metabolites often accumulate in specific plant organs in particular time of the vegetative stage of the plant. Some substances can only be isolated from extremely rare plants which is not a choice for sustainable production. Therefore, alternative production systems for plant-derived compounds are needed. The biotechnological production, that is, producing the plant secondary metabolites in cultured plant cells in large bioreactors may offer an attractive alternative approach.

Biotechnological Production Options

The production of a secondary metabolite of interest for industrial needs is often a challenge. As explained above, these compounds accumulate in plants in small quantities. The biotechnological production of high-value plant secondary metabolites therefore is a viable option to isolation processes from the intact plants or to the total chemical synthesis.

Biotechnology focuses on the exploitation of metabolic properties of living organisms for the production of valuable products of a very different structural and organizational level for the benefit of humans. The organisms vary from microbes (bacteria, fungi, yeast) to plants and animals. Over the decades, many laboratories all over the world have studied the possibilities to produce desired secondary metabolites using plant cell or tissue cultures. Cell cultures have been established from many plants, but often they do not produce sufficient amounts of the required secondary metabolites or the production is unstable. Various classical optimization tools have been applied (see in detail section "Enhancing the Production by Classical Optimization"), but very few success stories exist contrary to many good examples using microbial production systems.

Molecular biology of plants has emerged enormously during the past decades, but still the plant metabolic engineering has met only limited success, again in sharp contrast to microorganisms. This is due to our limited knowledge on complex biosynthesis of secondary metabolites. Despite the rapid development of not only plant genomics but also of analytical tools, genetic maps of biosynthetic pathways are far from complete. Furthermore, regulation of the individual steps leading to the desired end product is poorly understood (section "Metabolic Engineering").

Plant Cell Cultures Plant cell culture is a method where plant cells are cultivated under sterile conditions *in vitro*. Commonly, cell cultures are established from callus tissues by cultivating callus in liquid medium, and cell aggregates are broken by either mechanically or by orbital shaking in the cultivation vessel. Plant cells are biosynthetically totipotent, which means that each cell in culture retains its complete genetic information and thus is able to produce the same metabolites as the parent plant. Plant cell cultures have been extensively exploited for various biotechnological applications as an alternative to the traditional agricultural cultivation of plants. The use of cell culture systems offers advantages to produce metabolites in a controlled environment, independent of climatic conditions and under conditions in which the different production parameters can be optimized. Plant cell cultures can be categorized in two main classes, differentiated and undifferentiated cell cultures. The former consists of, for example, organs like shoots, roots, or embryos, whereas callus and cell suspension cultures are referred to as undifferentiated cell cultures. Since the first gene transfers in plants in 1983, achieved by four independently working groups [10–13], a number of efficient gene transfer techniques have been developed for genetic engineering of plants. In addition to so-called direct gene transfer techniques (e.g., particle bombardment, electroporation, microinjection), *Agrobacterium*-mediated gene transfer has been the most commonly used method for gene delivery to plants.

Hairy Root Cultures *Agrobacterium* (Rhizobiaceae) is a soil bacterium, which is able to deliver its own plasmid-DNA into the nuclear genome of the plant cell. The bacterium attaches into the wound site of the plant tissue and recognizes certain wound substances, for example, acetosyringone, secreted by the plant [14]. As a result, the *vir* (virulence) region of the plasmid becomes activated and processing of the T-DNA (transferred DNA) for the gene transfer starts [14, 15]. After successful integration of the bacterial DNA into the host plant genome, the tumor formation in the wound site begins as well as the production of low molecular weight tumor substances called opines. The opines are used as a nutrient for the bacterium [16]. The host range of *Agrobacterium* is perhaps broader than that of any other plant pathogenic bacterium, although a number of cultivated monocotyledonous plants and legumes are not natural hosts for this bacterium. The molecular mechanism of the resistance to *Agrobacterium* is not known, although the production of antimicrobial metabolites [17], a lack of *vir* gene inducers [18], inefficient T-DNA integration [19], and *Agrobacterium*-induced programmed cell death [20] have all been suggested. Successful gene transfer in monocot plants via *Agrobacterium* has been performed with maize, rice, wheat, and barley [21].

Hairy root is a plant disease caused by the infection of *Agrobacterium rhizogenes* carrying Ri (root-inducing) plasmid. During infection of the plant, the T-DNA of the Ri-plasmid is transferred and integrated in the nuclear genome of the host. As a result of the transformation, hairy roots appear at the infection site [22]. In the T-DNA, there are four genetic loci, called *rol*A, *rol*B, *rol*C, and *rol*D, which are responsible for the hairy root phenotype. These genes were shown to positively affect the secondary metabolite production in *Nicotiana* [23] and in *Atropa* [24]. Hairy roots are able to grow without externally supplied auxins, and certain *aux* genes from *Agrobacterium* have been shown to provide transformed cells with an additional source of auxin [25]. This is a clear advantage when considering the costs for large-scale cultivation. Hairy roots characteristically lack geotropism and have a high degree of lateral branching. In addition, hairy root cultures have demonstrated their ability to rapidly produce biomass as well as high contents of secondary metabolites, for example, tropane alkaloids [26, 27]. In Table 2, some pharmaceutical compounds produced by hairy

Medicinal Plants, Engineering of Secondary Metabolites in Cell Cultures. Table 2 Examples of metabolites produced by transformed hairy root cultures (adopted mainly from [28, 29])

Metabolite	Species	Activity	Reference
Ajmalicine, ajmaline	*Rauvolfia micrantha*	Antihypertensive	[30]
Artemisinin	*A. annua*	Antimalarial	[31]
Benzylisoquinoline alkaloids	*P. somniferum; E. californica*	Analgesic, antibiotic	[32]
Betalains	*Beta vulgaris*	Antioxidant, colorant	[33]
Camptothecin	*Ophiorrhiza pumila; Camptotheca acuminata*	Antitumor	[34, 35]
Iridoid glycosides	*Harpagophytum procumbens*	Anti-inflammatory, analgesic, and antidiabetic	[36]
3,4-Dihydroxy-L-phenylalanine	*Stizolobium hassjoo*	Therapeutic agent against Parkinson's disease	[37]
Rutin, hispidulin and syringin	*Saussurea involucrata*	Anti-inflammatory, antifungal	[38]
Scopolamine, hyoscyamine and atropine	*A. belladonna*	Anticholinergic	[24, 39]
Scopolamine and hyoscyamine	*Datura innoxia*	Anticholinergic	[40]
Scopolamine and hyoscyamine	*Datura quercifolia*	Anticholinergic	[41]
Scopolamine	*Duboisia leichhardtii*	Anticholinergic	[42]
Scopolamine and hyoscyamine	*Datura candida*	Anticholinergic	[43]
Scopolamine and hyoscyamine	*Datura innoxia*	Anticholinergic	[44]
Scopolamine and hyoscyamine	*H. niger*	Anticholinergic	[40]
Scopolamine and hyoscyamine	*H. muticus*	Anticholinergic	[26]
Scopolamine and hyoscyamine	*H. muticus, Nicotiana tabacum*	Anticholinergic	[45]
Scopolamine	*H. niger*	Anticholinergic	[46]
Solasodine	*Solanum khasianum*	Steroid hormone precursor	[47]
Paclitaxel	*Taxus brevifolia*	Anticancer	[48]
Terpenoid indole alkaloids	*C. roseus*	Antitumor	[49]
Thiarubrine A	*Ambrosia artemisiifolia*	Antifungal, antibacterail, antiviral	[50]
6-Methoxy-podophyllotoxin	*Linum album; Linum persicum*	Anticancer	[51]
Quinine, quinidine	*Cinchona ledgeriana*	Antimalarial	[52]
(+) catechin, (−) epicatechin-3-O-gallate, procyanidin B_2-3′-O-gallate	*Fagopyrun esculentum*	Antioxidant	[53]
Anthraquinone	*Rubia tinctoria*	Antimalarial, antineoplastic	[54]
Thiophene	*Tagetes patula*	Anti-inflammatory precursor	[55]
Valpotriates	*Valeriana officinalis*	Tranquilizing	[56]

root cultures are presented. Unlike crown gall tumors, hairy roots are capable of spontaneously regenerating into plants [57].

Bioreactors The selection of a suitable bioreactor type for the specific process depends on the desired product and the production material, for example, whether the production involves growing undifferentiated cells, hairy roots, or plantlets. Plants cells are larger in size than those of microbial cells, making them more sensitive to shear forces. For this reason, bioreactors have been designed where conventional mechanical impeller stirring have been replaced by bubble or wave-type agitation. Most widely used bioreactors are stirred tanks [58], but also airlift and bubble column reactors have been used in cultivation of plant cells. The classical production of shikonin is performed in airlift type of bioreactors. A balloon-type bubble bioreactor has been successfully used for the cultivation of, for example, ginseng roots [59].

One of the more recent developments in bioreactor design for plant cell applications has been the use of disposable bioreactors, usually plastic bags. Major advantages in these are that the capital costs are much lower than that of common stainless steel tanks. The production of glucocerobrosidase used for treating the enzyme deficiency cased in Gaucher's disease is performed in carrot cells grown in disposable bioreactors by Israeli company Protalix Biotherapeutics (www.protalix.com). The only secondary metabolite of pharmaceutical value, paclitaxel (Taxol®), is commercially produced in *Taxus* cells by German company Phyton Biotech (www.phytonbiotech.com). Moreover, lower expenses allow multiple parallel units to be employed, and high sterility requirements are met when there is no need for costly cleaning processes between runs. Disposable bioreactors may consist of a rigid cultivation container (tube, plate, flask, cylindrical vessel) or a flexible container (bag) [60]. Issues restricting the use of disposable bioreactors arise from a limited experience in their usage, insufficient strength of a plastic material, limited applicability of advanced automatization, and lack of suitable disposable sensors. Wave-mixed bioreactors [61], such as BioWave®, are well suited for small- to middle-scale processes for the production (Fig. 1) of, for example, plant-based secondary metabolites and therapeutic proteins, as well as cultivation of hairy roots [62, 63]. One of the highest productivities reported to date for paclitaxel production in *Taxus baccata* cell suspension cultures was achieved with immobilized cells cultivated in BioWave® system [64, 65].

Important factors when designing the cultivation of plant cell suspension cultures in bioreactors include guaranteed sterility through the whole process and low-shear mixing allowing still efficient nutrient transport without causing sedimentation or a loss in viability of the cells. In addition, the possibility for application of light induction for heterotrophic, photomixotrophic, and photoautotrophic cultures might be relevant [62]. Major physical process parameters regarding cultivation of plant cell and tissue cultures are temperature, viscosity, gas flow rates, and foaming.

Sometimes the lack of end-product formation may be due to the feedback inhibition, degradation of the product in the culture medium, or due to volatility of the substrates or end products. In such cases, adding of extra phase as a site for product accumulation might lead to increased production of the desired substance [66]. For example, addition of amberlite resin and charcoal resulted in increased accumulation of anthraquinones and vanilla, and coniferyl aldehyde, respectively [67–69]. On the other hand, bioconversion of water-insoluble substrates in cell culture systems can be aided by using cyclodextrins. They form inclusion bodies in their cyclodextrin cavity and by this way increase the water solubility of the substrates [70].

Enhancing the Production by Classical Optimization

Selection of High-Producing Lines

Selection of individual plants with desired traits has been a traditional approach in plant breeding. Similarly, high producers have been selected for further use, for example, for cloning and as a starting material for cell cultures. However, cell clones from the same origin may vary considerably in their metabolite production capacities. Selecting high producers is thus a very empirical approach, requiring a huge amount of screening work before good producing individuals are found [71, 72]. In order to obtain good producing cells, mutation strategies or application of various selective

Medicinal Plants, Engineering of Secondary Metabolites in Cell Cultures. Figure 1
Wave bioreactor is used to culture various types of plant cells. This is a 2-L disposable bag in a Wave® reactor containing tobacco hairy roots

agents, such as *p*-fluorophenylalanine [73], 5-methyltryptophan [74], or biotin [75], have been used. Although undifferentiated plant cells can be maintained in an undifferentiated state using phytohormones, they are not genetically or epigenetically stable. The concept of somaclonal variation was introduced by Larkin and Scowcroft in the beginning of 1980s, standing for the genetic variability in tissue culture–derived plants or cell culture clones [76]. These changes causing the variation can occur as large rearrangements in chromosomal level, for example, changes in chromosome number, karyotype modifications, or changes in gene level.

Somaclonal variation can be exploited when searching for high secondary metabolite producers or high producers of biomass, although a clear disadvantage is that these changes cannot be predicted or controlled and moreover, they are not always stable or heritable. The effect of culture age on growth rates were observed with *Nicotiana plumbaginifolia*, which showed higher growth rates with older cultures compared to newer cultures [77]. These differences were thought to appear as a cause of higher proportion of cells in older cultures exhibiting mutations which elevate cyclin-dependent kinases. Changes in ploidy levels are reported to affect regeneration capacity [78], gene silencing [79], and secondary metabolite production [80, 81]. After choosing good-producing cell lines, cultivation over time requires usually continuous selection in order to maintain high production levels. However, a gradual loss of secondary metabolite productivity over time is an obstacle in the development of commercial plant cell culture production systems [82, 83].

Optimization of Culture Medium

One of the major advantages in using plant cell cultures is the possibility of controlled and contained production systems. When attempting to reach high production levels, key roles are played by the composition of nutrient medium and other cultivation parameters, such as temperature, light, phytohormones, and gas exchange.

Because the plant cell is a production factory, the first requirement for obtaining high levels of products is the generation of high amounts of biomass or at least enough biomass for economic product yields. Plant cell cultures are usually grown heterotrophically using simple sugars as carbon source, sucrose being the most commonly used. Carbon source effects mainly on primary metabolism and by this way affects the overall productivity with either increased or decreased biomass production. Sucrose level may also have an indirect impact on secondary metabolite production, as inverse correlation between sucrose and hyoscyamine production was observed in *Hyoscyamus muticus* hairy root cultures [84]. This was probably due to the increased glycolysis and respiration rate with simultaneous overriding of secondary metabolite production. Sucrose is commonly applied in approximately 3% (w/v) concentration, but levels as high as 8% (w/v) have shown to increase the accumulation of indole and benzophenanthridine alkaloids in cell cultures of *Catharanthus roseus* and *Eschscholtzia californica*, respectively [85, 86].

Phosphate and nitrogen levels are perhaps the most important macronutrient factors effecting the secondary metabolite formation. Phosphate usually promotes cell growth, but often has been accompanied by lower secondary product formation. In fact, very often cell proliferation has been accompanied by decrease in secondary product formation and vice versa. For this reason, a two-stage cultivation system could be considered, where the cells are first cultivated in the medium optimized for cell multiplication and then transferred into medium limiting the biomass growth whereas enabling maximum product formation. As an example, shikonin produced by *Lithospermum erythrorhizon* in commercial scale by this type of two-phase system [87]. Low phosphate levels often have been correlated with high secondary metabolite formation, for example, in case of alkaloids in *Datura stramonium* [88], *Nicotiana tabacum* [89], and *C. roseus* [90]. Nitrogen is an important building block of amino acids, nucleic acids, proteins, and vitamins. Generally, nitrogen is added in the form of nitrate or ammonium, and the ratio of these salts plays an important role in secondary metabolite production of the plant cells. Reducing the levels of nitrogen generally leads to lower biomass production and thus leads to higher secondary metabolite production, as in the case of anthocyanin production by *Vitis vinifera* [91].

Phytohormones have an extensive effect not only on growth of plant cells, but also on differentiation and secondary metabolite production. Both the type and concentration of auxin and cytokinin as well as their ratio alter the growth and metabolite production dramatically in cultured plant cells. High auxin levels are known to inhibit the formation of secondary metabolites in a large number of cases, for example, tobacco alkaloids [92] with the simultaneous activation of polyamine conjugate biosynthesis [93]. Sometimes, replacement of synthetic auxin 2,4-D (2,4-dichlorophenoxy acetic acid) by NAA (naphthalene acetic acid) or natural auxin IAA (indole acetic acid) has shown to enhance the production of anthraquinones, shikonin, or anthocyanins [94–96].

Commonly understanding of cell culture behavior has been relied on the measurements of culture average parameters, such as cell density and metabolite profiles. However, due to the nature of plant cell division, in which daughter cells often remain attached through cell wall, aggregates of various sizes in cell suspension culture are formed. Thus, each aggregate is exposed to different microenvironmental conditions with respect to nutrient and oxygen availability between inner and outer regions of the aggregate [97]. Understanding such subpopulation dynamics and cellular variability using tools such as flow cytometry is important in order to control the culture as a whole.

Effect of Elicitors

The enhanced production of secondary metabolites from plant cell and tissue cultures through elicitation has opened up a new area of research which could have beneficial influences for pharmaceutical industry. Elicitors are compounds, biotic or abiotic, or even physical factors, which can trigger various defense-related reactions, and thereby induce secondary metabolite formation in plant cells. The mechanisms of how elicitors activate the respective genes and the whole biosynthetic machinery in a plant cell are under active investigation. However, it is evident that the gene expression occurs very quickly after the elicitor contact and many hours before the secondary metabolites are accumulated in a plant cell [98].

In general, elicitors can be categorized based on their molecular structure and origin. Biotic elicitors include compounds such as chitosan, alginate, pectin, chitin or they may contain complex mixtures of compounds like those of fungal or yeast extracts [99]. Abiotic elicitors are chemical compounds of nonbiological origin, for example, heavy metals and vanadate derivatives, or physical factors such as thermal or osmotic stress, UV-irradiation, or wounding. In particular, widely used elicitors for plant cell culture systems are jasmonates and jasmonic acid derivatives, which are naturally occurring hormones involved in the regulation of defence-related genes and act as signaling compounds in these reactions [100]. Application of jasmonates can result in large alterations in desired metabolites in *Catharanthus* [101, 102], in *Taxus* [103], and in *Nicotiana* [98]. Even though plant cells accumulate secondary metabolites typical for species in question independent of the type of elicitor used, the accumulation kinetics may vary greatly with different elicitors. Moreover, elicitors can effect on the release of desired secondary metabolite from the cell to the cultivation medium [104]. This is beneficial when considering the biotechnological production facilitating thus the downstream processing.

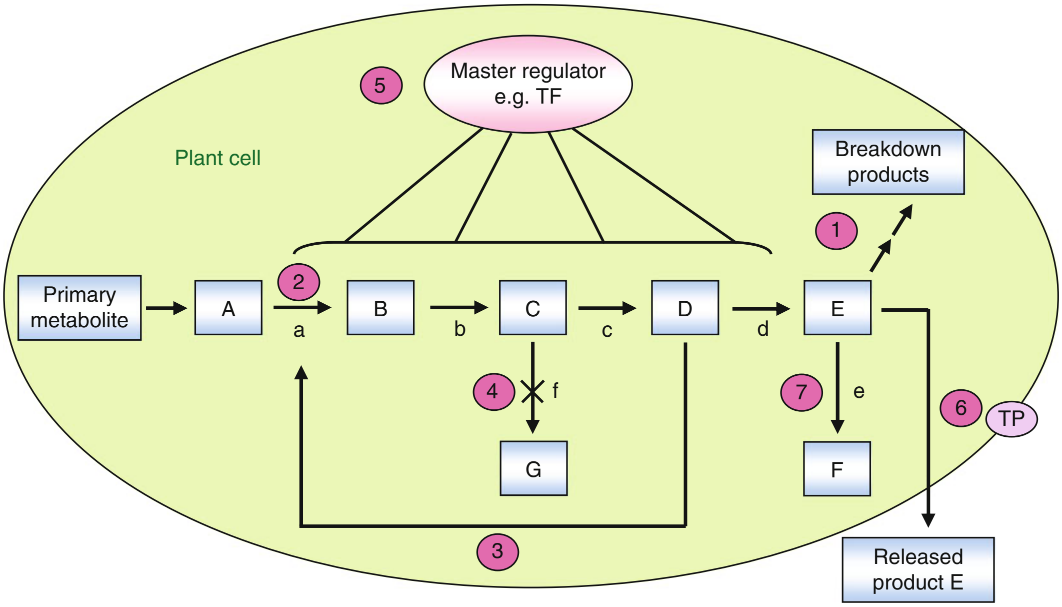
Generally, both the elicitor concentration and the length of elicitor application have to be determined for each cell culture individually [104]. Commonly it is thought that the best growth phase for the start of the elicitation is during the exponential growth phase when the enzymatic machinery for elicitor response is most active [105]. In addition, the composition of the culture medium, especially phytohormones, has a major impact on elicitor response. For example, divergent regulation by auxins on the biosynthesis of different metabolites in terpenoid indole alkaloid pathway was observed by *C. roseus* cell cultures [102]. This regulation by auxins was shown to be partly dependent on the presence of methyl jasmonate. Production of various plant-derived medicinal compounds has been successfully induced by using elicitors [106]. Unfortunately, many elicitors also cause a loss of viability of the producing cells, thus a thorough optimization of the whole production process is required when using elicitation.

Metabolic Engineering

Functional genomics tools offer now huge potential to engineer plant metabolic pathways toward the targeted end product or alternatively to form entirely novel structures through combinatorial biochemistry. However, rational engineering of secondary metabolite pathways requires a thorough knowledge of the whole biosynthetic pathway and detailed understanding of the regulatory mechanisms controlling the flux of the pathway (Fig. 2) [7]. Such information is not available for vast majority of secondary metabolites, explaining why only limited success has been obtained by metabolic engineering. New genome-wide transcript profiling techniques combined with up-to-date metabolomics allow us now to establish novel gene-to-gene and gene-to-metabolite networks which facilitate the gene discovery also in non-model plants that include most medicinal plants [102]. The ability to switch on entire pathways by ectopic expression of transcription factors suggests new possibilities for engineering secondary metabolite pathways (Fig. 2). Consequently, the utilization of plant cell cultures for biotechnological production of high-value alkaloids would thus become a true viable alternative.

Gene Discovery

Since the sequencing of *Arabidopsis* genome in 2,000 several other plants are being sequenced but still today very limited information exists for any medicinal plant. Therefore, also the biosynthetic pathways in these plants are largely unknown at the gene level. Several approaches have been developed to identify enzymes and the corresponding genes that are responsible for different biosynthetic pathway steps. One of the classical methods is the identification and isolation of intermediates and enzymes *via* precursor feeding [107]. The other very basic approach is to use cDNA libraries to identify genes by PCR amplification with primers designed to recognize conserved regions on the basis of enzyme homology from other plants with already known sequences [108]. More recently, methods based on differential display comparing mRNA transcripts of elicited and non-elicited cell culture samples have shown their potential in gene discovery. Goossens and coworkers [98] and Rischer and coworkers [102] utilized cDNA-AFLP technique for genome-wide gene hunt, whereas [109] supplemented their search with homology-based analysis of a cDNA library of elicited cells. In addition, the use of random sequencing of



Medicinal Plants, Engineering of Secondary Metabolites in Cell Cultures. Figure 2
The hypothetic scheme how the secondary metabolite E could be formed from primary metabolites via different enzymatic steps and how the biosynthesis could be regulated in a plant cell [7]. Engineering of secondary metabolite pathways is a series of complex events. The following strategies could be used to modify the production of hypothetical plant metabolite E: (1) decrease the catabolism of the desired compound, (2) enhance the expression/activity of a rate limiting enzyme, (3) prevent feedback inhibition of a key enzyme, (4) decrease the flux through competitive pathways, (5) enhance expression/activity of all genes involved in the pathway, (6) compartmentalize the desired compound, and (7) convert an existing product into a new product. *TF* transcription factor, *TP* transporter gene

elicited cDNA library can lead to identification of clones involved in the biosynthetic route in question as proven in case of *Taxus* biosynthesis [110].

The use of microarrays as widely used for model plants such as *Arabidopsis* is usually not applicable to medicinal plants simply because none has been sequenced with the very recent exception of tobacco http://www.pngg.org/tgi/index.html. The rapid advance of deep sequencing, however, will soon result in many important species being investigated at genome scale. The 454 pyrosequencing technique is currently perhaps the most widely used so-called next-generation sequencing technique for the *de novo* sequencing and analysis of transcriptomes in non-model organisms like medicinal plants are. For example, the GS FLX Titanium can generate one million reads with an average length of 400 bases at 99.5% accuracy. This technology was successfully used to discover putative genes involved in ginsenoside biosynthesis [111].

Once the candidate genes are discovered, they are functionally tested alone or in combination to find out their real mode of action, for example, improving or altering the production of desired metabolite. This is time consuming, and therefore new high-throughput systems have been developed, for example, miniaturized cell culture formats and multigene transformations.

Controlling the Expression of Transgenes

In order to be able to modify the metabolite profile of a respective medicinal plant or cell culture, the gene expression of target proteins and enzymes needs to be fine-tuned in an appropriate manner. For that purpose, the elements involved in transcriptional regulation of gene expression should be well characterized and evaluated to ensure correct spatial and temporal display. This also minimizes the potential adverse effects, and the outcome will be as wanted. Specific DNA sequences upstream of the encoding region of a gene that are recognized by proteins (transcription factors) involved in the initiation of transcription are determined as promoters. It is noteworthy that the promoter sequence itself is present in all tissues and cells, and thus the activity is controlled via transcription factors and their abundance. This opens the possibility to boost a cascade of enzymes and influence in the whole biosynthetic pathway in question by overexpressing transcription factors [112].

Promoters used for the metabolic engineering purposes can be divided into three classes:

1. Constitutive, that is, promoters that are continuously on in most or all of the tissues
2. Organ- or stage-specific, that is, promoters controlling spatiotemporal activity of the transgene
3. Inducible that are regulated by an external trigger of chemical or physical nature [113, 114].

As an example of the constitutive promoters and also the most used one in plant genetic engineering is the *Cauliflower mosaic virus 35S* promoter [115, 116]. The CaMV 35S promoter has been very thoroughly characterized and currently a typical CaMV 35S promoter in plant vectors consists of a bit more than one third of the full-length sequence [117]. It has also been observed that a partial duplication from −343 to −90 amplifies expression up to tenfold [118]. This promoter is also the most used one in metabolic engineering of plant cell cultures [119]. For the secondary metabolite production, the hairy root cultures have shown most potent, and little promise has been found with undifferentiated suspension cultures [120]. Actually there exist no studies for trying to find most suitable callus or suspension culture–specific promoters for efficient expression of target genes. This might be one factor why the success in using undifferentiated plant cell cultures for the production of valuable secondary metabolites has been so poor. However, the main blame for this is the current limited understanding of how the metabolic pathways and fluxes of secondary metabolites work in general.

Nowadays that the multigene transformations [121] are paving the way for more accurate and complex engineering of phenotypes, there is also more need to apply different promoter deployment strategies to reach the wanted goals. The delivery of 10–20 genes at the time is already very demanding, and thus there is no space for failure in running their expression. Roughly, two ways of proceeding can be drawn for promoter choice: utilization of the same promoter to run all the genes or combination of promoters to run different target genes in the generated multigene transformants. The use of same promoter carries the risk of triggering gene silencing. It is very important to increase the promoter diversity via promoter discovery and generation of synthetic sequences to run the expression. Perhaps one of the most interesting ways is to apply bidirectional sequences which allow simultaneous expression of two genes, and thus halves the number of required promoters for multigene engineering [122].

Targeting the Metabolic Enzymes

From the genetic engineering perspective of medicinal plants, one of the key elements is to express the genes in question in right tissues, and even more importantly target the respective enzymes to correct, specific subcellular compartments. A good example of compartmentalization is the biosynthesis of terpenoids that are synthesized from universal five-carbon precursors isopentenyl diphosphate (IPP) and dimethylallyl diphosphate (DMAPP), which in turn are formed via two alternate biosynthetic pathways localized in different subcellular compartments. The cytosolic mevalonic acid (MVA) pathway starts with condensation of two molecules of acetyl-CoA into acetoacetyl-CoA and finally gives rise to IPP. The methylerythritol phosphate (MEP) pathway takes place in plastids and leads to the formation of IPP and DMAPP from pyruvate and glyceraldehyde phosphate. The IPP and DMAPP precursors are then processed with prenyl diphosphate synthases in different compartments giving rise to intermediates that serve as substrates to a large group of terpene synthases resulting in construction of the final terpenoids [123, 124]. However, the picture is never black and white, and the subcellular localization studies as well as the genetic engineering experiments have shown that such a thing as a general rule does not apply to all tissues and species. From the rational genetic engineering point of view, this makes things even far more complex and we still need to reveal several aspects of biosynthetic pathways.

Targeting the biosynthetic enzymes to non-original compartment can also lead to interesting results. Precursors can be available in other compartments, and introduction of the respective enzyme can lead to increased accumulation of target compounds. For example, Wu and coworkers [125] showed that redirecting the sesquiterpene pathway from its natural cytosolic location to chloroplasts increased patchoulol accumulation even up to 10,000-fold when compared to native situation. Another example was given by introducing three different targeting modes: cytosolic, plastid, and ER of limonene synthase in transgenic tobacco plants [126]. Both the cytosolic and plastid targeting resulted in limonene formation, whereas ER targeting gave no response probably due to false folding or instability of the protein.

There has also been discussion on so-called metabolic channeling, which means that enzymes from the same pathway, especially the ones committing successive steps, form a protein complex resulting in efficient reactions and regulation of the pathway [127–129]. Aharoni and coworkers [130] interpreted that this might be a cause why some pathways do not seem to proceed even though substantial amount of substrate seem to be available. As a solution, an artificial channeling is suggested with the help of fusion constructs to be applied in the metabolic engineering. These studies also highlight the need for fluxomics and thorough understanding of metabolic pathways (see Sect. "Controlling the Expression of Transgenes").

Multigene Transformation

The first multigene-carrying transgenic plants were created either with several rounds of crossings between transgenic lines or by transforming transgenic plants with a new set of genes [131, 132]. The current multigene delivery systems are co-transformations with either linked or unlinked genes, that is, genes within a same vector or different vectors, respectively. The transfer itself is carried out either via *Agrobacterium*-mediated or direct transformation techniques. These systems have been developed mainly with crop plants, and the target pathways have been on nutritional composition like in engineering of the carotenoid pathway [133, 134]. These pioneer works have opened the possibility to engineer metabolic pathways of medicinal plants, and the potential in these can be seen almost as limitless. The future aim is the creation of a SMART locus (stable multiple arrays of transgenes), that is, a transgenic locus containing multiple genes, thus avoiding segregation in meiosis and possibly also minimizing rearrangements and silencing [121]. For medicinal plants, the possibility to modify entire metabolic pathways, to introduce completely new pathways, and to study complex metabolic control circuits and regulations are perhaps the main future goals.

New Compounds by Engineered Enzymes/Proteins

In most common approaches, the intention of metabolic engineering is to either overexpress or repress genes leading to the accumulation of certain compounds (Fig. 2). The first successful genetic engineering approach to the medicinal plant was performed already almost 20 years ago. Yun and coworkers [135] introduced the gene-encoding hyoscyamine-6β-hydroxylase (H6H) from *Hyoscyamus niger* to the medicinal plant *A. belladonna*. As a result of the overexpression of *h6h*, the plants produced almost exclusively scopolamine, whereas in the control plants the production of hyoscyamine (precursor of scopolamine) was dominant. Later, the function of the same gene was demonstrated to be different in hairy roots of *Hyoscyamus muticus* [26]. The overexpression of *h6h* caused 100-fold increase in scopolamine production, whereas the hyoscyamine contents were not reduced.

There are also examples where genetic engineering can lead to formation of entirely new metabolites. Classically, this can, for example, be achieved by generating somatic hybrids, that is, by exposing enzymes and regulators derived from different genomes to new environments. A good example is the production of demissidine in somatic *Solanum* hybrids neither parent of which contained this specific metabolite but only a set of different precursors [136].

More recently, the combinatorial biochemistry concept which is based on the fact that enzymes often show relaxed substrate specificity, that is, that they can under certain conditions process substrates which differ from the preferred one is exploited in a stricter sense. Usually, native genes are modified with the aim of creating modified enzymes catalyzing new reactions. Initially, attempts to alter the substrate specificity of plant-derived terpenoid synthases by rather unspecific methods such as mutagenesis or truncation were quite unpredictable [137]. Meanwhile, however, it could be shown that preselection of a mutant strictosidine synthase with a specific point mutation according to substrate acceptance results in quite predictable events. *C. roseus* hairy roots expressing the gene formed unnatural terpenoid indole alkaloids when were fed with derivatized precursors in contrast to the wild type [138].

Future Directions

Different omics in techniques have opened totally new avenues to discover genes, to learn about their functions, for example, transcription, and to finally map the biosynthetic pathways leading to the formation of important secondary metabolites. Metabolomics, which deals with all cellular metabolites, was first defined in microbiology but has also been recognized as an important sector of post-genome plant science [139]. Even in the absence of any visible change in a cell or individual plant, metabolomics, which allows phenotyping by exhaustive metabolic profiling, can show how cells respond as a system. Plant metabolomics is of particular importance because of the huge chemical diversity in plants compared to microorganisms and animals [140]. The number of metabolites from the plant kingdom has been estimated at 200,000 or even more [6], and each plant has its own complex set of metabolites. By integrating transcriptome and metabolome data, one can build networks and get insight on how particular metabolites are formed in plants [102, 140]. This in turn helps us to identify the key genes that could be engineered for the production of improved medicinal plants.

Since cell physiology involves dynamic rather than static processes, the investigation of fluxes is needed to complement phenotyping by metabolomics which only allows inventory, although time-resolved snapshots. However, in contrast to mammalian and microbial cells, flux quantification in plants is much less advanced. This is mainly due to the high degree of subcellular compartmentation and the complexity which arises from intercompartmental transport. Labeling experiments have been very successfully used already in the past for the elucidation of biosynthetic pathways in plants [141], but flux determination has only recently gained pace due to the fast development of analytical and computational technology. Analytical techniques of choice are nuclear magnetic resonance (NMR) spectrometry and mass spectrometry (MS) [142]. Generally, there are two fundamentally different methods available facilitating flux measurement – steady-state and dynamic analysis – both of which have certain restrictions and benefits [143]. The latter, that is, kinetic approach is particularly interesting in the sense that it potentially could lead to predictive modeling in regard to secondary metabolism, while steady-state analysis is mainly used to measure carbon flux in well-defined pathways of primary metabolism [144].

In conclusion, modern genomic tools allow for mass gene discovery from plants although many biosynthetic pathways are incompletely resolved and medicinal plants have rarely been sequenced. Nevertheless, predictive metabolic engineering remains a goal of the future. This is because transgene integration in higher plants occurs through illegitimate rather than homologous recombination. DNA integration is random with a preference for gene-rich regions. Gene disruptions, sequence changes, and the production of new proteins constitute common consequences resulting in either predictable or unpredictable effects [145]. In this situation, the power of functional genomics tools allowing the comprehensive investigation of biological systems cannot be overemphasized. Genomics identifies all genes of a plant, while transcriptomics and proteomics provide information about their activities in cells or organs under certain conditions, and finally metabolomics and fluxomics account for the accumulation and kinetics of metabolites, that is, the phenotype. The individual techniques as such are thus invaluable to assign functions, but the real advantage lays in their combination, that is, the systems biology approach [140]. Interestingly at the same time, these tools allow not only the investigation of artificial situations generated by man but also for the first time broad assessment of natural variation.

Acknowledgments

This work has been financially supported by the SmartCell project (nr. 222176) from European Commission Framework 7 programme.

Bibliography

Primary Literature

1. Hostettmann K, Terreaux C (2000) Search for new lead compounds from higher plants. Chimia 54:652–657
2. Müller-Kuhrt L (2003) Putting nature back into drug discovery. Nat Biotechnol 21:602
3. Newman DJ, Cragg GM (2007) Natural products as sources of new drugs over the last 25 years. J Nat Prod 70: 461–477
4. Verpoorte R (1998) Exploration of nature's chemodiversity: the role of secondary metabolites as leads in drug development. Drug Discov Today 3:232–238
5. De Luca V, St Pierre B (2000) The cell and developmental biology of alkaloid biosynthesis. Trends Plant Sci 5: 168–173
6. Hartmann T, Kutchan TM, Strack D (2005) Evolution of metabolic diversity. Phytochemistry 66:1198–1199
7. Oksman-Caldentey K-M, Inzé D (2004) Plant cell factories in the post-genomic era: new ways to produce designer secondary metabolites. Trends Plant Sci 9:433–440
8. Verpoorte R (2000) Secondary metabolism. In: Verpoorte R, Alfermann AW (eds) Metabolic engineering of plant secondary metabolism. Kluwer, Dordrech, pp 1–29
9. Ziegler J, Facchini PJ (2008) Alkaloid biosynthesis – metabolism and trafficking. Annu Rev Plant Biol 59:735–769
10. Bevan MW, Flavell RB, Chilton MD (1983) A chimaeric antibiotic resistance gene as a selectable marker for plant cell transformation. Nature 304:184–187
11. Fraley RT, Rogers SG, Horsch RB, Sanders PR, Flick JS, Adams SP, Bittner ML, Brand LA, Fink CL, Fry JS, Galluppi GR, Goldberg SB, Hoffman NL, Woo SC (1983) Expression of bacterial genes in plant cells. Proc Natl Acad Sci 80:4803–4807
12. Herrera-Estrella L, Depicker M, van Montagu M, Schell J (1983) Expression of chimaeric genes transferred into plant cells using a Ti-plasmid-derived vector. Nature 303:209–213
13. Murai N, Sutton DW, Murray MG, Slightom JL, Merlo DJ, Reichert NA, Sengupta-Gopalan C, Stock CA, Barker RF, Kemp JD, Hall TC (1983) Phaseolin gene from bean is expressed after transfer to sunflower via tumor-inducing plasmid vectors. Science 222:476–482
14. Zupan J, Muth TR, Draper O, Zambryski P (2000) The transfer of DNA from *Agrobacterium tumefaciens* into plants: a feast of fundamental insights. Plant J 23:11–28
15. Sheng J, Citovsky V (1996) *Agrobacterium* – plant cell DNA transport: have virulence proteins, will travel. Plant Cell 8:1699–1710
16. Chilton M-D, Tepfer DA, Petit A, David C, Casse-Delbart T, Tempé J (1982) *Agrobacterium rhizogenes* inserts T-DNA into the genomes of the host plant root cells. Nature 295:432–434
17. Sahi SV, Chilton M-D, Chilton WS (1990) Corn metabolites affect growth and virulence of *Agrobacterium tumefaciens*. Proc Natl Acad Sci USA 87:3879–3883
18. Usami S, Morikawa S, Takebe I, Machida Y (1987) Absence in monocotyledonous plants at the diffusible plant factors inducing T-DNA circularization and *vir* gene expression in *Agrobacterium*. Mol Gen Genet 209:221–226
19. Narasimhulu SB, Deng X, Sarria R, Gelvin SB (1996) Early transcription of *Agrobacterium* T-DNA genes in tobacco and maize. Plant Cell 8:873–886
20. Hansen G (2000) Evidence for *Agrobacterium*-induced apoptosis in maize cells. Mol Plant Microbe Interact 13: 649–657
21. Nadolska-Orczyk A, Orczyk W, Przetakiewicz A (2000) *Agrobacterium* -mediated transformation of cereals – from technique development to its application. Acta Physiol Plant 22:77–88
22. Sevón N, Oksman-Caldentey K-M (2002) *Agrobacterium rhizogenes*-mediated transformation: root cultures as a source of alkaloids. Planta Med 68:859–868
23. Palazón J, Cusidó RM, Roig C, Piñol MT (1998) Expression of the *rol*C gene and nicotine production in transgenic roots and their regenerated plants. Plant Cell Rep 17:384–390
24. Bonhomme V, Laurain-Mattar D, Lacoux J, Fliniaux M-A, Jacquin-Dubreuil A (2000) Tropane alkaloid production by hairy roots of *Atropa belladonna* obtained after transformation with *Agrobacterium rhizogenes* 15834 and *Agrobacterium tumefaciens* containing *rol A*, *B*, *C* genes only. J Biotechnol 81:151–158
25. Chriqui D, Guivarch A, Dewitte W, Prinsen E, van Onkelen H (1996) *Rol* genes and root initiation and development. Plant Soil 187:47–55
26. Jouhikainen K, Lindgren L, Jokelainen T, Hiltunen R, Teeri T, Oksman-Caldentey K-M (1999) Enhancement of scopolamine production in *Hyoscyamus muticus* L. hairy root cultures by genetic engineering. Planta 208:545–551
27. Zhang L, Ding R, Chai Y, Bonfill M, Moyano E, Oksman-Caldentey K-M, Xu T, Pi Y, Wang Z, Zhang H, Kai G, Liao Z, Sun K, Tang K (2004) Engineering tropane alkaloid pathway in *Hyoscyamus niger* hairy root cultures. Proc Natl Acad Sci USA 101:6786–6791. doi:6786
28. Georgiev MI, Pavlov AI, Bley T (2007) Hairy root type plant in vitro systems as sources of bioactive substances. Appl Microbiol Biotechnol 74:1175–1185
29. Srivastava S, Srivastava AK (2007) Hairy root culture for mass-production of high-value secondary metabolites. Crit Rev Biotechnol 27:29–43
30. Sudha CG, Obul RB, Ravishankar GA, Seeni S (2003) Production of ajmalicine and ajmaline in hairy root cultures of *Rauvolfia micrantha* Hook F., a rare and endemic medicinal plant. Biotechnol Lett 25:631–636
31. Weathers P, Bunk G, McCoy MC (2005) The effect of phytohormones on growth and artemisinin production in *Artemisia annua* hairy roots. In Vitro Cell Dev B 41:47–53
32. Park S-U, Facchini P (2000) *Agrobacterium rhizogenes*-mediated transformation of opium poppy, *Papaver somniferum* L., and California poppy, *Eschscholzia californica* Cham., root cultures. J Exp Bot 347:1005–1006

33. Pavlov A, Bley T (2006) Betalains biosynthesis by *Beta vulgaris* L. hairy root culture in different bioreactor systems. Process Biochem 41:848–852
34. Saito K, Sudo H, Yamazaki M, Koseki-Nakamura M, Kitajima M, Takayama H, Aimi N (2001) Feasible production of camptothecin by hairy root culture of *Ophiorrhiza pumila*. Plant Cell Rep 20:267–271
35. Lorence A, Medina-Bolivar F, Nessler CL (2004) Camptothecin and 10-hydroxycamptothecin from *Camptotheca acuminata* hairy roots. Plant Cell Rep 22:437–441
36. Georgiev M, Heinrich M, Kerns G, Pavlov A, Bley T (2006) Production of iridoids and phenolics by transformed *Harpagophytum procumbens* root cultures. Eng Life Sci 6:593–596
37. Sung H, Huang S-Y (2006) Medium optimization of transformed root cultures of *Stizolobium hassjoo* producing L-DOPA with response surface methodology. Biotechnol Bioeng 94:441–447
38. Fu C-X, Xu Y-J, Zhao D-X, Ma FS (2006) A comparison between hairy root cultures and wild plants of *Saussurea involucrata* in phenylpropanoids production. Plant Cell Rep 24:750–754
39. Jung G, Tepfer D (1987) Use of genetic transformation by the Ri T-DNA of *Agrobacterium rhizogenes* to stimulate biomass and tropane alkaloid production in *Atropa belladonna* and *Calystegia sepium* roots grown *in vitro*. J Ferment Bioeng 85:454–457
40. Shimomura K, Sauerwein M, Ishimaru K (1991) Tropane alkaloids in the adventitious and hairy root cultures of *Solanaceous* plants. Phytochemistry 30:2275–2278
41. Dupraz JM, Christen P, Kapetanidis I (1994) Tropane alkaloids in transformed roots of *Datura quercifolia*. Planta Med 60: 158–162
42. Mano Y, Ohkawa H, Yamada Y (1989) Production of tropane alkaloids by hairy root cultures of *Duboisia Leichhardtii* transformed by *Agrobacterium rhizogenes*. Plant Sci 59: 191–201
43. Christen P, Robert MF, Phillipson JD, Evans WC (1991) Alkaloids of hairy root cultures of a *Datura candida* hybrid. Plant Cell Rep 9:101–104
44. Dechaux C, Boitel-Conti M (2005) A Strategy for overaccumulation of scopolamine in *Datura innoxia* hairy root cultures. Acta Biol Cracov Bot 47:101–107
45. Häkkinen ST, Moyano E, Cusidó RM, Palazón J, Piñol MT, Oksman-Caldentey K-M (2005) Enhanced secretion of tropane alkaloids in *Nicotiana tabacum* hairy roots expressing heterologous hyoscyamine-6beta-hydroxylase. J Exp Bot 420: 2611–2618
46. Zhang L, Ding R, Chai Y, Bonfill M, Moyano E, Oksman-Caldentey K-M, Xu T, Pi Y, Wang Z, Zhang H, Kai G, Liao Z, Sun X, Tang K (2004) Engineering tropane biosynthetic pathway in *Hyoscyamus niger* hairy root cultures. Proc Natl Acad Sci 1117 USA 101:6786–6791
47. Jacob A, Malpathak N (2004) Green hairy root cultures of *Solanum khasianum* Clarke – a new route to *in vitro* solasodine production. Curr Sci 87:1442–1447
48. Huang Z, Mu Y, Zhou Y, Chen W, Xu K, Yu Z, Bian Y, Yang Q (1997) Transformation of *Taxus brevifolia* by *Agrobacterium rhizogenes* and taxol production in hairy root culture. Acta Bot Yunnanica 19:292–296
49. Palazón J, Cusidó RM, Gonzalo J, Bonfill M, Morales C, Piñol MT (1998) Relation between the amount of rolC gene product and indole alkaloid accumulation in *Catharanthus roseus* transformed root cultures. J Plant Physiol 153:712–718
50. Bhagwath SG, Hjortso MA (2000) Statistical analysis of elicitation strategies for thiarubrine. A production in hairy root cultures of *Ambrosia artemisiifolia*. J Biotechnol 80: 159–167
51. Wink M, Alfermann AW, Franke R, Wetterauer B, Distl M, Windhoevel J, Krohn O, Fuss E, Garden H, Mohagheghzadeh A, Wildi EJ, Ripplinger P (2005) Sustainable bioproduction of phytochemicals by plant *in vitro* cultures: anticancer agents. Plant Gene Res 3:90–100
52. Hamill JD, Robins RJ, Rhodes MJC (1989) Alkaloid production by transformed root cultures of *Cinchona ledgeriana*. Planta Med 55:354–357
53. Trotin F, Moumou Y, Vasseur J (1993) Flavanol production by *Fagopyrum esculentum* hairy and normal root cultures. Phytochemistry 32:929–931
54. Sato K, Yamazaki T, Okuyama E, Yoshihira K & Shimomura K (1991) Anthraquinones production by transformed root cultures of *Rubia tinctorum*: Influence of phytohormones and sucrose concentration. Phytochemistry 30:1507–1509
55. Croes AF, Vander Berg AJR, Bosveld M, Breteler H, Wullems GJ (1989) Thiophene accumulation in relation to morphology in roots of *Tagetes patula*. Effects of auxin and transformation by *Agrobacterium*. Planta Med 179:43–50
56. Granicher F, Christen P, Kapetandis I (1992) High-yield production of valepotriates by hairy root cultures of *Valeriana officnalis* L. var. *sambucifolia* Mikan. Plant Cell Rep 11:339–342
57. Oksman-Caldentey K-M, Kivelä O, Hiltunen R (1991) Spontaneous shoot organogenesis and plant regeneration from hairy root cultures of *Hyoscyamus muticus*. Plant Sci 78:129–136
58. Su WW (2006) Bioreactor engineering for recombinant protein production using plant cell suspension culture. In: Gupta SD, Ibaraki Y (eds) Plant tissue culture engineering. Springer, The Netherlands, pp 135–159
59. Choi YE, Kim YS, Paek KY (2006) Types and designs of bioreactors for hairy root culture. In: Gupta SD, Ibaraki Y (eds) Plant tissue culture engineering. Springer, The Netherlands, pp 161–172
60. Eibl R, Kaiser S, Lombriser R, Eibl D (2010) Disposable bioreactors: the current state-of-art and recommended applications in biotechnology. Appl Microbiol Biotechnol 86:41–49
61. Eibl R, Werner S, Eibl D (2009) Bag bioreactor based on wave-induced motion: characteristics and applications. Adv Biochem Eng Biotechnol 115:55–87
62. Eibl R, Eibl D (2002) Bioreactors for plant cell and tissue cultures. In: Oksman-Caldentey K-M, Barz WH (eds) Plant biotechnology and transgenic plants. Marcel Dekker, Basel, pp 163–199

63. Palazón J, Mallol A, Lettenbauer C, Cusidó RM, Piñol MT (2003) Growth and gingenoside production in hairy root cultures of *Panax ginseng* using a novel bioreactor. Planta Med 69:344–349
64. Bentebibel S, Moyano E, Palazón J, Cusidó RM, Bonfill M, Eibl R, Piñol MT (2005) Effects of immobilization by entrapment in alginate and scale-up on paclitaxel and baccatin III production in cell suspension cultures of *Taxus baccata*. Biotechnol Bioeng 89:647–655
65. Bonfill M, Bentebibel S, Moyano E, Palazón J, Cusidó RM, Piñol MT (2008) Paclitaxel and baccatin III production induced by methyl jasmonate in free and immobilized cells of *Taxus baccata*. Biol Plant 51:647–652
66. Ramachandra Rao S, Ravishankar GA (2002) Plant cell cultures: chemical factories of secondary metabolites. Biotechnol Adv 20:101–153
67. Robins RJ, Rhodes MJC (1986) The stimulation of anthraquinone production by *Cinchona ledgeriana* cultures with polymeric adsorbents. Appl Microbiol Biotechnol 24:35–41
68. Knuth ME, Sahai OP (1991) Flavour composition and method. US Patent 5,068,184, 26 Nov 1991
69. Beiderbeck R, Knoop B (1987) Two-phase culture. In: Constael F, Vasil I (eds) Cell culture and somatic cell genetics of plants, vol 5. Academic, San Diego, pp 255–266
70. Van Uden W, Woedenbag HJ, Pras N (1994) Cyclodextrins as a useful tool for bioconversion in plant cell biotechnology. Plant Cell Tiss Org 38:103–113
71. Oksman-Caldentey K-M, Vuorela H, Strauss A, Hiltunen R (1987) Variation in the tropane alkaloid content of *Hyoscyamus muticus* plants and cell culture clones. Planta Med 53:349–354
72. Mano Y, Ohkawa H, Yamada Y (1989) Production of tropane alkaloids by hairy root cultures of *Duboisia leichhardtii* transformed by *Agrobacterium rhizogenes*. Plant Sci 59:191–201
73. Berlin J (1980) Para-fluorophenylalanine resistant cell lines of tobacco. Z Pflanzenphysiol 97:317–324
74. Widholm JM (1974) Evidence for compartmentation of tryptophan in cultured plant tissues. Free tryptophan levels and inhibition of anthranilate synthetase. Physiol Plant 30:323–326
75. Wataneba K, Yano SI, Yamada Y (1982) Selection of cultured plant cell lines producing high levels of biotin. Phytochemicals 21:513–516
76. Larkin PJ, Scowcroft WR (1981) Somaclonal variation – a novel source of variability from cell cultures for plant improvement. Theor Appl Genet 60:197–214
77. Zhang KR, John PCL (2005) Raised level of cyclin dependent kinase A after prolonged suspension culture of *Nicotiana plumbaginifolia* is associated with more rapid growth and division, diminished cytoskeleton and lost capacity for regeneration: implications for instability of cultures plant cells. Plant Cell Tissue Organ Cult 82:295–308
78. Shiba T, Mii M (2005) Visual selection and maintenance of the cell lines with high plant regeneration ability and low ploidy level in *Dianthus acicularis* by monitoring with flow cytometry analysis. Plant Cell Rep 24:572–580
79. Pikaard CS (2001) Genomic change and gene silencing in polyploids. Trends Genet 17:675–677
80. Hirasuna TJ, Pestchanker LJ, Srinivasan V, Shuler ML (1996) Taxol production in suspension cultures of *Taxus baccata*. Plant Cell Tiss Org 44:95–102
81. Wallaart TE, Pras N, Quax WJ (1999) Seasonal variations of artemisinin and its biosynthetic precursors in tetraploid *Artemisia annua* plants compared with the diploid wild-type. Planta Med 65:723–728
82. Deus-Neumann B, Zenk MH (1984) Instability of indole alkaloid production in *Catharanthus roseus* cell suspension-cultures. Planta Med 50:427–431
83. Qu JG, Zhang W, Yu XJ, Jin MF (2005) Instability of anthocyanin accumulation in *Vitis vinifera* L. var. Gamay Freaux suspension cultures. Biotechnol Bioprocess Eng 10:155–161
84. Wilhelmson A, Häkkinen ST, Kallio P, Oksman-Caldentey K-M, Nuutila AM (2006) Heterologous expression of *Vitreoscilla* hemoglobin (VHb) and cultivation conditions affect the alkaloid profile of *Hyoscyamus muticus* hairy roots. Biotechnol Prog 22:350–358
85. Knobloch KH, Berlin J (1980) Influence of medium composition on the formation of secondary compounds in cell suspension cultures of *Catharanthus roseus* L. G. Don. Z Naturforsch 35C:551–556
86. Berlin J, Forche E, Wray V, Hammer J, Hosel W (1983) Formation of benzophenanthridine alkaloids by suspension cultures of *Eschscholtzia californica*. Z Naturforsch 38:346–352
87. Fujita Y, Tabata M, Nishi A, Yamada Y (1982) New medium and production of secondary compounds with two-staged culture medium. In: Fujiwara A (ed) Plant tissue culture. Maruzen, Tokyo, pp 399–400
88. Payne J, Hamill JD, Robins RJ, Rhodes MJC (1987) Production of hyoscyamine by "hairy root" cultures of *Datura stramonium*. Planta Med 53:474–478
89. Mantell SH, Pearson DW, Hazell LP, Smith H (1983) The effect of initial phosphate and sucrose levels on nicotine accumulation in batch suspension cultures of *Nicotiana tabacum* L. Plant Cell Rep 1:73–77
90. Toivonen L, Ojala M, Kauppinen V (1991) Studies on the optimization of growth and indole alkalooid production by hairy root cultures of *Catharanthus roseus*. Biotechnol Bioeng 37:673–680
91. Do CB, Cormier F (1991) Effects of low nitrate and high sugar concentrations on anthocyanin content and composition of grape (*Vitis vinifera* L.) cell suspension. Plant Cell Rep 9:500–504
92. Ishikawa A, Yoshihara T, Nakamura K (1994) Jasmonate-inducible expression of a potato cathepsin D inhibitor-GUS gene fusion in tobacco cells. Plant Mol Biol 26:403–414
93. Tiburcio AF, Kaur-Sawhney R, Ingersoll R, Galston AW (1985) Correlation between polyamines and pyrrolidine in developing tobacco callus. Plant Physiol 78:323–326
94. Zenk MH, El-Shagi, E, Schulte U (1975) Anthraquinone production by cell suspension cultures of *Morinda citrifolia*. Planta Med 28:79–101

95. Tabata M (1988) Naphtoquinones. In: Constael F, Vasil I (eds) Cell culture and somatic cell genetics of plants, vol 5. Academic, San Diego, pp 99–111

96. Rajendran L, Ravishankar GA, Venkataraman LV, Prathiba KR (1992) Anthocyanin production in callus cultures of *Daucus carota* L. as influenced by nutrient stress and osmoticum. Biotechnol Lett 14:707–714

97. Kolewe ME, Gaurav V, Roberst SC (2008) Pharmaceutically active natural product synthesis and supply via plant cell culture technology. Mol Pharm 5:243–256

98. Goossens A, Häkkinen ST, Laakso I, Seppänen-Laakso T, Biondi S, De Sutter V, Lammertyn F, Nuutila AM, Söderlund H, Zabeau M, Inzé D, Oksman-Caldentey K-M (2003) A functional genomics approach toward the understanding of secondary metabolism in plant cells. Proc Natl Acad Sci USA 100:8595–8600

99. Vasconsuelo AA, Boland R (2007) Molecular aspects of the early stages of elicitation of secondary metabolites in plants. Plant Sci 172:861–875

100. Pauwels L, Barbero GF, Geerinck J, Tilleman S, Grunewald W, Pérez AC, Chico JM, Bossche RV, Sewell J, Gil E, García-Casado G, Witters E, Inzé D, Long JA, De Jaeger G, Solano R, Goossens A (2010) NINJA connects the co-repressor TOPLESS to jasmonate signalling. Nature 464:788–791

101. Lee-Parsons CW, Royce AJ (2006) Precursor limitations in methyl jasmonate-induced *Catharanthus roseus* cell cultures. Plant Cell Rep 25:607–612

102. Rischer H, Oreŝič M, Seppänen-Laakso T, Katajamaa M, Lammertyn F, Ardiles-Diaz W, Van Montagu MCE, Inzé D, Oksman-Caldentey K-M, Goossens A (2006) Gene-to-metabolite networks for terpenoid indole alkaloid biosynthesis in *Catharanthus roseus* cells. Proc Natl Acad Sci 103:5614–5619

103. Yukimune Y, Tabata H, Higashi Y, Hara Y (1996) Methyl jasmonate-induced overproduction of paclitaxel and baccatin III in *Taxus* cell suspension cultures. Nat Biotechnol 14:1129–1132

104. Sevón N, Hiltunen R, Oksman-Caldentey K-M (1992) Chitosan increases hyoscyamine content in hairy root cultures of *Hyoscyamus muticus*. Pharm Pharmacol Lett 2:96–99

105. Vasconsuelo AA, Giuletti AM, Picotto G, Rodriguez-Talou J, Boland R (2003) Involvement of the PLC/PKC pathway in chitosan-induced anthraquinone production by *Rubia tinctorium* L. cell cultures. Plant Sci 165:429–436

106. Namdeo AG (2010) Plant cell elicitation for production of secondary metabolites: a review. Pharmacogn Rev 1:69–79

107. Bringmann G, Wohlfarth M, Rischer H, Grüne M, Schlauer J (2000) A new biosynthetic pathway to alkaloids in plants: acetogenic isoquinolines. Angew Chem Int Ed 39:1464–1466

108. Wildung MR, Croteau R (1996) A cDNA clone for taxadiene synthase, the diterpene synthase, the diterpene cyclise that catalyzes the committed step of taxol biosynthesis. J Biol Chem 271:9201–9204

109. Kaspera R, Croteau R (2006) Cytochrome P450 oxygenases of taxol biosynthesis. Phytochem Rev 5:433–444

110. Jennewein S, Wildung MR, Chau M, Walker K, Croteau R (2004) Random sequencing of an induced *Taxus* cell cDNA library for identification of clones involved in taxol biosynthesis. Proc Natl Acad Sci USA 101:9149–9154

111. Sun C, Li Y, Wu Q, Luo H, Sun Y, Song J, Lui EMK, Chen S (2010) *De novo* sequencing and analysis of the American ginseng root transcriptome using a GS FLX titanium platform to discover putative genes involved in ginsenoside biosynthesis. BMC Genomics 11:262–273

112. Memelink J, Verpoorte R, Kijne JW (2001) ORCAnization of jasmonate-responsive gene expression in alkaloid metabolism. Trends Plant Sci 6:212–219

113. Potenza C, Aleman L, Sengupta-Gopalan C (2004) Targeting transgene expression in research, agricultural, and environmental applications:promoters used in plant transformation. In Vitro Cell Dev B 40:1–22

114. Yoshida K, Shinmyo A (2000) Transgene expression systems in plant, a natural bioreactor. J Biosci Bioeng 90:353–362

115. Guilley H, Dudley RK, Jonard G, Balázs E, Richards KE (1982) Transcription of cauliflower mosaic virus DNA: detection of promoter sequences, and characterization of transcripts. Cell 30:763–773

116. Odell JT, Nagy F, Chua NH (1985) Identification of DNA sequences required for activity of the cauliflower mosaic virus 35S promoter. Nature 313:810–812

117. Fang RX, Nagy F, Sivasubramanian S, Chua NH (1989) Multiple *cis* regulatory elements for maximal expression of the cauliflower mosaic virus 35S promoter in transgenic plants. Plant Cell 1:141–150

118. Kay R, Chan A, Daly M, McPherson J (1987) Duplication of CaMV 35S promoter sequences creates a strong enhancer for plant genes. Science 236:1299–1302

119. Zárate R, Verpoorte R (2007) Strategies for the genetic modification of the medicinal plant *Catharanthus roseus* (L.) G. Don. Phytochem Rev 6:475–491

120. Weathers PJ, Towler MJ, Xu J (2010) Bench to batch: advances in plant cell culture for producing useful products. Appl Microbiol Biotechnol 85:1339–1351

121. Naqvi S, Farré G, Sanahuja G, Capell T, Zhu C, Christou P (2010) When more is better: multigene engineering in plants. Trends Plant Sci 15:48–56

122. Peremarti A, Twyman RM, Gómes-Galera S, Naqvi S, Farré G, Sabalza M, Miralpeix B, Dashevskaya S, Yuan D, Ramessar K, Christou P, Zhu C, Bassie L, Capell T (2010) Promoter diversity in multigene transformation. Plant Mol Biol 73:363–378

123. Dudareva N, Negre F, Nagegowda DA, Orlova I (2006) Plant volatiles: recent advances and future prospects. Crit Rev Plant Sci 25:417–440

124. Nagegowda DA (2010) Plant volatile terpenoid metabolism: biosynthetic genes, transcriptional regulation and subcellular compartmentation. FEBS Lett 584:2965–2973

125. Wu S, Schalk M, Clark A, Miles RB, Coates R, Chappell J (2006) Redirection of cytosolic or plastidic isoprenoid precursors elevates terpene production in plants. Nat Biotechnol 24:1441–1447

126. Ohara K, Ujihara T, Endo T, Sato F, Yazaki K (2003) Limonene production in tobacco with *Perilla* limonene synthase cDNA. J Exp Bot 54:2635–2642
127. Chappell J, Wolf F, Proulx J, Cuella R, Saunders C (1995) Is the reaction catalyzed by 3-hydroxy-3-methylglutaryl coenzyme A reductase a rate-limiting step for isoprenoid biosynthesis in plants. Plant Physiol 109:1337–1343
128. Winkel BSJ (2004) Metabolic channeling in plants. Annu Rev Plant Biol 55:85–107
129. Kristensen C, Morant M, Olsen CE, Ekstrøm CT, Galbraith DW, Lindberg Møller B, Bak S (2005) Metabolic engineering of dhurrin in transgenic *Arabidopsis* plants with marginal inadvertent effects on the metabolome and transcriptome. Proc Natl Acad Sci USA 102:1779–1784
130. Aharoni A, Jongsma MA, Bouwmeester HJ (2005) Volatile science? Metabolic engineering of terpenoids in plants. Trends Plant Sci 10:594–602
131. Ma JK, Hiatt A, Hein M, Vine ND, Wang F, Stabila P, van Dolleweerd C, Mostov K, Lehner T (1995) Generation and assembly of secretory antibodies in plants. Science 268:716–719
132. Jobling SA, Westcott RJ, Tayal A, Jeffcoat R, Schwall GP (2002) Production of a freeze-thaw-stable potato starch by antisense inhibition of three starch synthase genes. Nat Biotechnol 20:295–299
133. Zhu C, Naqvi S, Breitenbach J, Sandmann G, Christou P, Capell T (2008) Combinatorial genetic transformation generates a library of metabolic phenotypes for the carotenoid pathway in maize. Proc Natl Acad Sci USA 105:18232–18237
134. Fujisawa M, Takita E, Harada H, Sakurai N, Suzuki H, Ohyama K, Shibata D, Misawa N (2009) Pathway engineering of *Brassica napus* seeds using multiple key enzyme genes involved in ketocarotenoid formation. J Exp Bot 60:1319–1332
135. Yun D-J, Hashimoto T, Yamada Y (1992) Metabolic engineering of medicinal plants: transgenic *Atropa belladonna* with an improved alkaloid composition. Proc Natl Acad Sci USA 89:11799–11803
136. Laurila J, Laakso I, Valkonen JPT, Hiltunen R, Pehu E (1996) Formation of parental-type and novel glycoalkaloids in somatic hybrids between *Solanum brevidens* and *S. tuberosum*. Plant Sci 118:145–155
137. Little DB, Croteau RB (2002) Alteration of product formation by directed mutagenesis and truncation of the multiple-product sesquiterpene synthases δ-selinene synthase and γ-humulene synthase. Arch Biochem Biophys 402:120–135
138. Runguphan W, O'Connor SE (2009) Metabolic reprogramming of periwinkle plant culture. Nat Chem Biol 5:151–153
139. Trethewey RN, Krozky AJ, Willmitzer L (1999) Metabolic profiling: a Rosetta stone for genomics? Curr Opin Plant Biol 2:83–85
140. Oksman-Caldentey K-M, Saito K (2005) Integrating genomics and metabolomics for engineering plant metabolic pathways. Curr Opin Biotechnol 16:174–179
141. Wheeler GL, Jones MA, Smirnoff N (1998) The biosynthetic pathway of vitamin C in higher plants. Nature 393:365–369
142. Ratcliffe RG, Shachar-Hill Y (2005) Revealing metabolic phenotypes in plants: inputs from NMR analysis. Biol Rev 80:27–43
143. Kruger NJ, Ratcliffe RG (2007) Dynamic metabolic networks: going with the flow. Phytochemistry 68:2136–2138
144. Kruger NJ, Huddleston JE, Le Lay P, Brown ND, Ratcliffe RG (2007) Network flux analysis: impact of 13C-substrates on metabolism in *Arabidopsis thaliana* cell suspension cultures. Phytochemistry 68:2176–2188
145. Rischer H, Oksman-Caldentey K-M (2006) Unintended effects in genetically modified crops: revealed by metabolomics? Trends Biotechnol 24:102–104

Books and Reviews

Allen DK, Libourel IG, Shachar-Hill Y (2009) Metabolic flux analysis in plants: coping with complexity. Plant Cell Environ 32:1241–1257

Bhagwath SG, Hjortso MA (2000) J Biotechnol 80:159–167

Bonhomme V, Laurain-Mattar D, Lacoux J, Fliniaux M, Jacquin-Dubreuil A (2000) J Biotechnol 81:151–158

Buchanan BB, Gruissem W, Russell LJ (eds) (2000) Biochemistry & molecular biology of plants. American Society of Plant Physiologists, Rockville, p 1367

Christen P, Robert MF, Phillipson JD, Evans WC (1991) Plant Cell Rep 9:101–104

Croes AF, Vander Berg AJR, Bosveld M, Breteler H, Wullems GJ (1989) Planta Med 179:43–50

Dechaux C, Boitel-Conti M (2005) Acta Biol Cracov Bot 47:101–107

Du H, Huang Y, Tang Y (2010) Genetic and metabolic engineering of isoflavonoid biosynthesis. Appl Microbiol Biotechnol 86:1293–1312

Dudareva N, Pichersky E (2008) Metabolic engineering of plant volatiles. Curr Opin Biotechnol 19:181–189

Dupraz JM, Christen P, Kapetanidis I (1994) Planta Med 60:158–162

Fu C-X, Xu Y-J, Zhao D-X, Ma FS (2006) Plant Cell Rep 24:750–754

Georgiev M, Heinrich M, Kerns G, Pavlov A, Bley T (2006) Eng Life Sci 6:593–596

Georgiev MI, Pavlov AI, Bley T (2007) Hairy root type plant in vitro systems as sources of bioactive substances. Appl Microbiol Biotechnol 74:1175–1185

Granicher F, Christen P, Kapetandis I (1992) Plant Cell Rep 11:339–342

Häkkinen ST, Moyano E, Cusidó RM, Palazón J, Piñol MT, Oksman-Caldentey K-M (2005) J Exp Bot 420:2611–2618

Häkkinen ST, Oksman-Caldentey K-M (2004) Regulation of secondary metabolism in tobacco cell cultures. In: Nagata T, Hasezawa S, Inzé D (eds) Biotechnology in agriculture and forestry, vol 53, Tobacco BY-2 cells. Springer, Berlin/Heidelberg, pp 231–249

Hamill JD, Robins RJ, Rhodes MJC (1989) Planta Med 55:354–357

Huang Z, Mu Y, Zhou Y, Chen W, Xu K, Yu Z, Bian Y, Yang Q (1997) Acta Bot Yunnanica 19:292–296

Jacob A, Malpathak N (2004) Curr Sci 87:1442–1447

Jouhikainen K, Lindgren L, Jokelainen T, Hiltunen R, Oksman-Caldentey K-M (1999) Enhancement of scopolamine production in *Hyscyamus muticus* L. hairy root cultures by genetic engineering. Planta 208:545–551

Jung G, Tepfer D (1987) J Ferment Bioeng 85:454–457

Lorence A, Medina-Bolivar F, Nessler CL (2004) Plant Cell Rep 22:437–441

Mano Y, Ohkawa H, Yamada Y (1989) Plant Sci 59:191–201

Nascimiento NC, Fett-Neto AG (2010) Plant secondary metabolism and challenges in modifying its operation: an overview. Meth Mol Biol 643:1–13

Oksman-Caldentey K-M, Barz W (eds) (2002) Plant biotechnology and transgenic plants. Marcel and Dekker, New York, p 719

Oksman-Caldentey K-M, Inzé D (2004) Plant cell factories in the post-genomic era: new ways to produce designer secondary metabolites. Trends Plant Sci 9:433–440

Oksman-Caldentey K-M, Inzé D, Orešič M (2004) Connecting genes to metabolites by a systems biology approach. Proc Natl Acad Sci USA 101:9949–9950

Palazón J, Cusidó RM, Gonzalo J, Bonfill M, Morales C, Piñol MT (1998) J Plant Physiol 153:712–718

Park S-U, Facchini P (2000) J Exp Bot 347:1005–1006

Pavlov A, Bley T (2006) Process Biochem 41:848–852

Rischer H, Oksman-Caldentey K-M (2005) Biotechnological utilization of plant genetic resources for the production of phytopharmaceuticals. Plant Gen Resour 3:83–89

Saito K, Dixon RD, Willmitzer L (2006) Plant metabolomics. In: Nagata T, Lörz H, Widholm JM (eds) Biotechnology in agriculture and forestry, vol 57. Springer, Berlin/Heidelberg, p 347

Saito K, Sudo H, Yamazaki M, Koseki-Nakamura M, Kitajima M, Takayama H, Aimi N (2001) Plant Cell Rep 20:267–271

Saito K, Yamazaki T, Okuyama E, Yoshihira K, Shimomura K (1991) Phytochemistry 30:2977–2980

Samuelsson G (2004) Drugs of natural origin. A textbook of pharmacognocy, 5th edn. Swedish Pharmaceutical Press, Stockholm, pp 473–575

Schäfer H, Wink M (2009) Medicinally important secondary metabolites in recombinant microorganisms or plants: progress in alkaloid biosynthesis. Biotechnol J 4:1684–1703

Sevón N, Oksman-Caldentey K-M (2002) *Agrobacterium rhizogenes*-mediated transformation: root cultures as a source of alkaloids. Planta Med 68:859–868

Shimomura K, Sauerwein M, Ishimaru K (1991) Phytochemistry 30:2275–2278

Srivastava S, Srivastava AK (2007) Hairy root culture for mass-production of high-value secondary metabolites. Crit Rev Biotechnol 27:29–43

Sudha CG, Obul RB, Ravishankar GA, Seeni S (2003) Biotechnol Lett 25:631–636

Sung H, Huang S-Y (2006) Biotechnol Bioeng 94:441–447

Trotin F, Moumou Y, Vasseur J (1993) Phytochemistry 32:929–931

Verpoorte R, Alfermann AW (eds) (2000) Metabolic engineering of plant secondary metabolism. Academic, Dordrecht, p 286

Verpoorte R, Alfermann AW, Johnson TS (eds) (2007) Applications of plant metabolic engineering. Springer, Dordrecht, p 332

Weathers P, Bunk G, McCoy MC (2005) In Vitro Cell Dev B 41:47–53

Wink M, Alfermann AW, Franke R, Wetterauer B, Distl M, Windhoevel J, Krohn O, Fuss E, Garden H, Mohagheghzadeh A, Wildi EJ, Ripplinger P (2005) Plant Gene Res 3:90–100

Zhang L, Ding R, Chai Y, Bonfill M, Moyano E, Oksman-Caldentey K-M, Xu T, Pi Y, Wang Z, Zhang H, Kai G, Liao Z, Sun X, Tang K (2004) Proc Natl Acad Sci USA 101:6786–6791

Membrane Electrolytes, from Perfluoro Sulfonic Acid (PFSA) to Hydrocarbon Ionomers

Kenji Miyatake
Clean Energy Research Center, University of Yamanashi, Kofu, Japan

Article Outline

Glossary

Hydrocarbon ionomers Polymers with hydrocarbon backbone (and generally no fluorine groups) and acidic functions.

Hydrophilic/hydrophobic phase separation Phase-separated morphology of ionomer membranes based on the differences in the hydrophilicity (and hydrophobicity) of the components.

Ion exchange capacity Amount of acidic or ion-exchangeable sites per weight or volume unit of ionomer membranes, often abbreviated as IEC. EW (equivalent weight, or weight of ionomer membranes per acidic or ion-exchangeable site) is a reciprocal of IEC.

Ionomers Originally defined as copolymers having one ionic group per polymer repeating unit, where composition of ion-containing copolymer unit is less than 20%. Nowadays, often used to be synonymous with polymer electrolytes.

Ionic channels Network of acidic groups and water molecules, through which proton and/or hydronium ions can migrate.

Perfluoro sulfonic acid ionomers Copolymers composed of poly(tetrafluoroethylene) and poly (trifluoroethylene) with perfluoro sulfonic acid ether side chains.

Proton exchange membranes Membranes that can exchange protons with other cations or that can transport protons.

Definition of the Subject and Its Importance

Proton exchange membranes (PEMs) are one of the key materials in low-temperature fuel cells; proton exchange membrane fuel cells (PEMFCs); and direct methanol fuel cells (DMFCs). Especially, recent trend in the research and development of low-temperature fuel cells focuses on PEMFCs for transportation (electric vehicle) applications due to the impact on economy and environment. The most important role of PEMs is to transport protons formed as a product of oxidation reaction of fuels at the anode to the cathode, where oxygen reduction reaction takes place to produce water. In addition to this, there are a number of requirements for PEM materials for the practical fuel cell applications, which include

1. Proton conductivity (higher than 0.01 S/cm, hopefully 0.1 S/cm)
2. Chemical, physical (mechanical and dimensional), and thermal stability
3. Impermeability of fuels (hydrogen, methanol) and oxidants (air, oxygen)
4. Water transport capability (high water flux) from the cathode to the anode

   These properties have to be assured under a wide range of temperature and humidity (−30–120°C, nominal 0–100% relative humidity (RH)) considering the fabrication of membrane electrode assemblies (MEAs)
5. Easy processability and compatibility with the electrodes are also crucial factors

   For wide dissemination of fuel cells
6. Environmental adaptability (recyclability or disposability)
7. Low cost (final target for electric vehicle applications would be cheaper than US\$ 10/m^2) need to be taken into account. While a number of PEMs have been developed, no single membranes fulfill all of these requirements. Currently, the most promising PEMs are perfluoro sulfonic acid (PFSA) ionomers. Another candidate second to the PFSA ionomers is non-fluorinated (or in some cases only slightly fluorinated) hydrocarbon ionomers. The aim of this chapter is to review the most recent progress on these two classes of ionomer membranes for low-temperature fuel cell applications

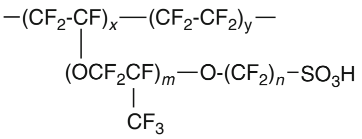
US DOE (Department of Energy) has set technical targets in PEMs for transportation applications (Table 1) [1]. The targets are for gas crossover (permeability), area-specific resistance, operating temperature, cost, and durability. In 2015, car companies will make a decision whether they continue their endeavor to commercialize fuel cell vehicles. The membrane scientists are facing a big challenge in order to help them go further with fuel cells.

Introduction

PFSA ionomers are copolymers of poly(tetrafluoroethylene) (PTFE) and poly(trifluoroethylene) with pendant perfluoro sulfonic acid groups. There are several industrial companies that supply the PFSA ionomers as resins, membranes, or solution. Such companies include DuPont, 3M, Asahi Kasei, Asahi Chemical, and Solvay Solexis. A general chemical structure of the PFSA ionomers is shown in Fig. 1. They share the similar chemical structure and the differences in them lie in the copolymer composition and the length of the pendant side chains and/or the presence of trifluoromethyl groups. Originally, DuPont developed the PFSAs in the 1950s. The PFSA membranes have a history as separator membranes in the chlor-alkali electrolysis industry. In this application, the PFSA membranes are used as single ion (Na$^+$) conductor. Typical PFSA membranes survive electrolysis operation under strongly basic conditions (>30% NaOH aq.) for more than several years retaining high current efficiency (>95%). Due to the hydrophobicity of fluorinated polymer main chains and strong acidity of the flexible side chains, the PFSA ionomer membranes show distinct hydrophilic/hydrophobic phase

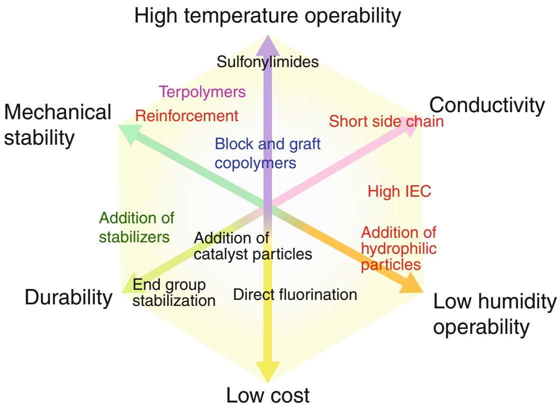
Membrane Electrolytes, from Perfluoro Sulfonic Acid (PFSA) to Hydrocarbon Ionomers. Table 1 US Department of Energy (DOE) targets in PEMs for transportation applications

Characteristics	Units	Year	
		2010	2015
O_2 crossover	mA/cm^2	2	2
H_2 crossover	mA/cm^2	2	2
Area-specific resistance at			
Maximum operating temperature and water partial pressures from 40 to 80 kPa	ohm cm^2	0.02	0.02
80°C water and water vapor partial pressure from 25 to 45 kPa	ohm cm^2	0.02	0.02
30°C water and water vapor partial pressure up to 4 kPa	ohm cm^2	0.03	0.03
−20°C	ohm cm^2	0.2	0.2
Maximum operating temperature	°C	120	120
Unassisted start from low temperature	°C	−40	−40
Cost	\$/m^2	20	20
Durability with cycling			
Mechanical	Cycles with <10 sccm crossover	20,000	20,000
Chemical	mA/cm^2 (H_2 crossover)	200	20



Membrane Electrolytes, from Perfluoro Sulfonic Acid (PFSA) to Hydrocarbon Ionomers. Figure 1
General chemical structure of the perfluoro sulfonic acid (PFSA) ionomers

separation. The sulfonic acid groups aggregate to form hydrophilic domains, while fluorinated main chains form hydrophobic domains with some crystallinity. When hydrated, hydrophilic domains (5–6 nm in diameter) contain water molecules and become interconnected, which are responsible for high proton transport capability or high proton conductivity. The morphology of PFSA membranes has been analyzed by small-angle X-ray scattering (SAXS), X-ray diffraction (XRD), small-angle neutron scattering (SANS), transmission electron microscopy (TEM), atomic force microscopy (AFM), and differential scanning calorimetry (DSC) techniques and well-reviewed in the books and articles. Short side chain and thus, high ion exchange capacity (IEC) and highly proton conductive PFSAs were developed by Dow and the membranes were used by Ballard Power Systems to realize high-performance PEMFCs in the mid-1980s. Since then, the PFSAs have been and will be the most studied PEMs for fuel cells. The recent research trends in PFSA membranes are summarized in Fig. 2. In order to fulfill the above-mentioned requirements, a considerable effort has been consumed especially in the last decade. They can be classified into six items; high-temperature operability, low-humidity operability, proton conductivity, mechanical stability, durability, and cost. Details of each approach can be found in the literature, and because of the limited space, only some of the representative examples are described in the next section.



Membrane Electrolytes, from Perfluoro Sulfonic Acid (PFSA) to Hydrocarbon Ionomers. Figure 2
Recent research trends in perfluoro sulfonic acid (PFSA) membranes

Hydrocarbon ionomers have a rather longer history (ca. 100 years) as cation exchange resins. Original hydrocarbon ionomers were based on the sulfonated polystyrenes or phenol resins. In the earliest stage of the PEMFC research, such ionomer membranes were investigated. GE invented PEMFC with hydrocarbon ionomer membranes in the early 1960s. The operation time of the initial PEMFCs was very limited by the membrane durability. PFSA membranes have replaced them in the mid-1960s, and since then they have been the main option. However, the hydrocarbon ionomer materials have been reexamined in more detail in the last decade due to their possible lower production cost, more freedom in molecular design and chemical modification, and better environmental compatibility compared with PFSA ionomers. Hydrocarbon ionomers can be roughly classified into two classes, aromatic and aliphatic ionomers depending on their main chain structure. Most effort has been focused on aromatic ionomers due to their chemical robustness. A number of so-called engineering plastics, such as poly(phenylene)s, polyimides, poly(arylene ether)s, and poly(ether ether ketones)s have been utilized as a base skeleton. Nevertheless, none of the existing hydrocarbon ionomer membranes can compete with the PFSA ionomer membranes. The most critical issues of hydrocarbon ionomer membranes are still insufficient durability and large dependence of the proton conductivity upon humidity. The challenge is to achieve these two conflicting properties within a single ionomer membrane. Recent effective approaches are reviewed below.

Perfluoro Sulfonic Acid Ionomer Membranes

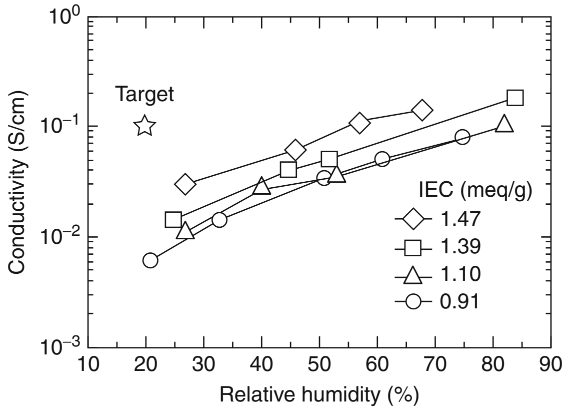
Short Side Chain PFSAs

The simplest way to improve the proton conductivity of ionomer membranes is to increase IEC, either by using monomers with short side chains or by increasing copolymer composition of sulfonic acid-containing units. The former approach seems preferable in terms of the mechanical properties of the membranes. This is the case of Dow membranes, which contain oxytetrafluoroethylene sulfonic acid groups. Due to the synthetic difficulties involving many reaction steps, Dow gave up supplying their short side chain PFSAs. However, due to their high potential as high proton conductive PEMs, a few kinds of short side chain PFSAs are currently available. In Table 2 are summarized chemical structure and molecular weight of representative monomers. Solvay Solexis has developed a simple preparation method of the short side chain vinyl monomers, which enabled them to produce the short side chain PFSAs at the industrial scale [2]. Nowadays, a number of short side chain PFSAs have been produced at the pilot scale from several companies [3, 4].

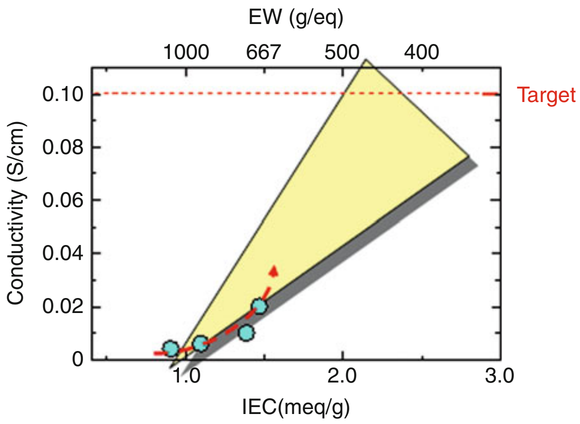
Figure 3 shows humidity dependence of the proton conductivity of PFSA membranes with different IEC values. In Fig. 4 is plotted proton conductivity of PFSA membranes at 110°C, 20% RH as a function of IEC value. The data shown in both figures clearly demonstrate that the proton conductivity increases with increasing IEC of the membrane material. However, there is still a gap in the conductivity between the current level and the target. The question is how high IEC would be required to reach 0.1 S/cm. Extrapolating the current line gives a rough value of IEC, 2–3 meq/g. Such high IEC membranes probably suffer from low mechanical and dimensional stability, which has to be addressed by appropriate molecular modifications such as cross-linking. Reinforcing with compatible resins (e.g., porous substrates or fabrics of PTFEs) may also be a possible option.

Membrane Electrolytes, from Perfluoro Sulfonic Acid (PFSA) to Hydrocarbon Ionomers. Table 2 Structure and molecular weight of monomers for PFSAs

Structure	Molecular weight (g/mol)	Suppliers
$CF_2 = CF - OCF_2\underset{\substack{\mid \\ CF_3}}{CF} - OCF_2CF_2 - SO_2F$	446	Du Pont, Asahi Chemical, Asahi Glass, etc.
$CF_2 = CF - O(CF_2CF_2)_2 - SO_2F$	380	3M, Asahi Kasei, etc.
$CF_2 = CF - OCF_2CF_2 - SO_2F$	280	Dow (currently not available), Solvay Solexis, etc.

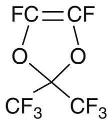


Membrane Electrolytes, from Perfluoro Sulfonic Acid (PFSA) to Hydrocarbon Ionomers. Figure 3
Humidity dependence of the proton conductivity of perfluoro sulfonic acid (PFSA) membranes at 110°C



Membrane Electrolytes, from Perfluoro Sulfonic Acid (PFSA) to Hydrocarbon Ionomers. Figure 4
Proton conductivity of perfluoro sulfonic acid (PFSA) membranes at 110°C and 20% RH

In addition to the proton-conducting properties, short side chain provides additional advantages, such as high glass transition temperature and high decomposition temperature. For example, Solvay Solexis's short side chain PFSAs, known as Aquivion, show α relaxation transition at 160°C, which is ca. 50°C higher than that of the conventional long side chain PFSAs. This is a consequence of higher crystallinity of the short side chain PFSAs [5, 6]. Therefore, higher IEC membranes can be prepared from the short side chain



Membrane Electrolytes, from Perfluoro Sulfonic Acid (PFSA) to Hydrocarbon Ionomers. Figure 5
Chemical structure of bulky comonomer for the perfluoro sulfonic acid (PFSA) terpolymers

PFSAs with the same crystallinity (e.g., similar thermal and mechanical properties) with the long side chain PFSAs, or the same IEC membranes with the higher crystallinity.

Terpolymers

Asahi Glass has investigated the effect of third compounds and found that cyclic vinyl compounds, for example, 2,2-bis(trifluoromethyl)-4,5-difluoro-1,3-dioxole, functions well to improve the thermal and dimensional stability (Fig. 5) [7]. Similar to the short side chain ionomers, the terpolymer PFSAs showed higher elastic modulus and ca. 40°C higher softening temperature than those of the conventional long side chain PFSAs. The water absorbability of the terpolymers was nearly half in a wide range of temperature.

There seems to be a number of other candidates suitable as a third comonomer (e.g., trifluorovinyl monomers containing phosphonic acid, sulfonamide, sulfonimide, and/or cross-linkable moieties) [8–12], however, high production cost of these rather complex perfluorinated vinyl compounds may be an obstacle to the practical industrial applications.

Stabilized PFSAs

Despite the established durability of PFSAs as separator membranes in the chlor-alkali electrolysis, they fail to function as proton exchange membranes for longtime fuel cells operation. Typical PFSA membranes degrade and have pinholes within several hundred hours of operation under high temperature, low humidity, and open circuit voltage (OCV) conditions. It is generally recognized that the degradation of PFSA membranes is caused by hydroxide (HO·) radical, which is generated by the homolysis or the Fenton's reaction (catalyzed by Fe^{2+} ions) of hydrogen peroxide. When oxygen reduction reaction (ORR) occurs via four-electron process, the product is water. Instead, two-electron process provides hydrogen peroxide as the product, which may occasionally form both at the cathode and the anode. At the anode where oxygen permeated through the membrane can react with hydrogen at the lower potential than that of the cathode, the ORR reaction is more likely to proceed via two-electron process. Hydroperoxide (HOO·) radical would also form in the operating fuel cells and cause membrane degradation similarly.

The probable degradation mechanisms (Fig. 6) of the PFSA membranes involving the radicals are

1. Decomposition of the end groups [13–15]

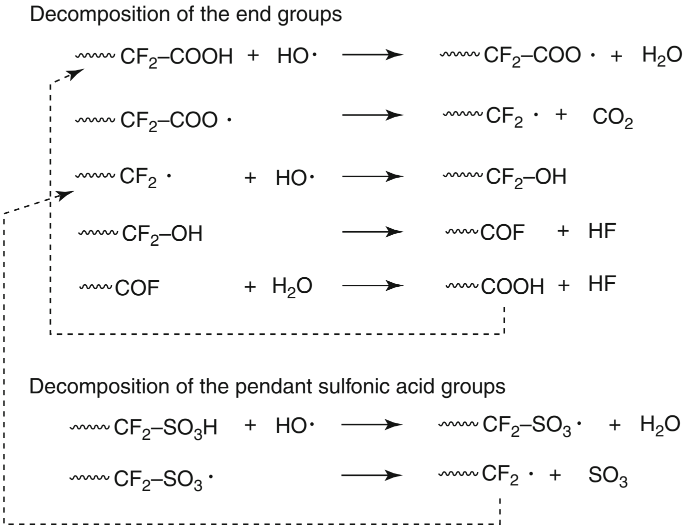
   PFSAs contain carboxylic acid groups at the end of polymer main chains as a result of the initiator in the polymerization reaction. The terminal carboxylic acid groups can be attacked by the radical species to produce shortened carboxylic acid groups and hydrogen fluoride. The repetition of this reaction causes unzipping degradation of polymer main chains.
2. Decomposition of the pendant sulfonic acid groups [16–18].

   In the side chains of PFSA ionomers, the sulfonic acid groups are likely to be attacked by the radicals. Under dry or low humidity conditions, in particular, most sulfonic acid groups are not dissociated and prone to hydrogen abstraction reaction by the radicals. The reaction gives $–CF_2^{\cdot}$ radicals and initiates the unzipping degradation of polymer main chains.

In addition to these mechanisms, there may be another degradation mechanism, in which fluorocarbons ($–CF_2–$) are hydrogenated to $–CH_2–$ with hydrogen gas and then attacked by radicals. It is shown by solid-state NMR spectroscopy that the side chain degradation is severer than the main chain degradation for the stabilized PFSAs.

It has been proposed that the terminal carboxylic acid (–COOH) groups can be converted to trifluoromethyl ($–CF_3$) groups by treating the ionomers with fluorine gas under heated conditions [15]. The degradation could be mitigated significantly with the trifluoromethylated ionomers. However, there remains ca. 10% of the degradation even extraporating the content of terminal carboxylic acid groups to zero. The remaining degradation is considered to be the side chain degradation.

While some transition metal ions promote radical formation and ionomer degradation, other transition metal ions act as radical quencher and mitigate the ionomer degradation (Fig. 7) [16, 17, 19–21]. Such ions include Ce^{3+} and Mn^{2+}, having reduction potential at 1.74 and 1.51 V at 25°C, respectively. These ions can reduce hydroxide radicals or hydrogen peroxide to


Decomposition of the end groups
$CF_2–COOH + HO\cdot \longrightarrow CF_2–COO\cdot + H_2O$
$CF_2–COO\cdot \longrightarrow CF_2\cdot + CO_2$
$CF_2\cdot + HO\cdot \longrightarrow CF_2–OH$
$CF_2–OH \longrightarrow COF + HF$
$COF + H_2O \longrightarrow COOH + HF$
Decomposition of the pendant sulfonic acid groups
$CF_2–SO_3H + HO\cdot \longrightarrow CF_2–SO_3\cdot + H_2O$
$CF_2–SO_3\cdot \longrightarrow CF_2\cdot + SO_3$

Membrane Electrolytes, from Perfluoro Sulfonic Acid (PFSA) to Hydrocarbon Ionomers. Figure 6
Possible degradation mechanisms of perfluoro sulfonic acid (PFSAs) involving oxidative radical species

water, but not reduce hydrogen peroxide to hydroxide radicals. The oxidized ions, Ce^{4+} and Mn^{3+}, could be reduced to Ce^{3+} and Mn^{2+} by hydrogen permeating through the ionomer membranes from the anode to the cathode. Typically, a few to 10% of sulfonic acid groups are ion-exchanged to such transition metal ions in order to obtain improved durability and acceptable proton conductivity. Some membranes have been claimed to survive OCV test at 120°C and 18% RH and constant current (0.2 A/cm^2) operation at 120°C and 50% RH for several thousand hours. In addition to the role as effective mitigants of chemical degradation, the multivalent transition metal ions function as ionic cross-linkers to render better mechanical properties to the ionomer membranes.

New Synthetic Routes

Synthetic approaches of sulfonated perfluorovinyl monomers are limited and require many steps. Fluorinated cyclic compounds (sultones and epoxides), which are often used as raw materials, need careful attention in handling. Such synthetic complexity is one of the reasons of their high production cost. A unique methodology to produce sulfonated perfluorovinyl monomers has been developed by Asahi Glass (Fig. 8) [7, 22]. The synthetic process, as they name PERFECT (PERFluorination of an Esterified Compound then Thermal elimination) process, involves perfluorination of partially fluorinated aliphatic esters containing sulfonyl fluoride groups. The perfluorination reaction with fluorine (F_2) is carried out in solution. The thermal decomposition of the perfluorinated esters gives acetyl fluorides, which can be converted to the corresponding vinyl monomers via conventional reaction pathway. This process may be applicable to a variety of aliphatic esters to provide new vinyl monomers with fewer steps and at lower production cost.

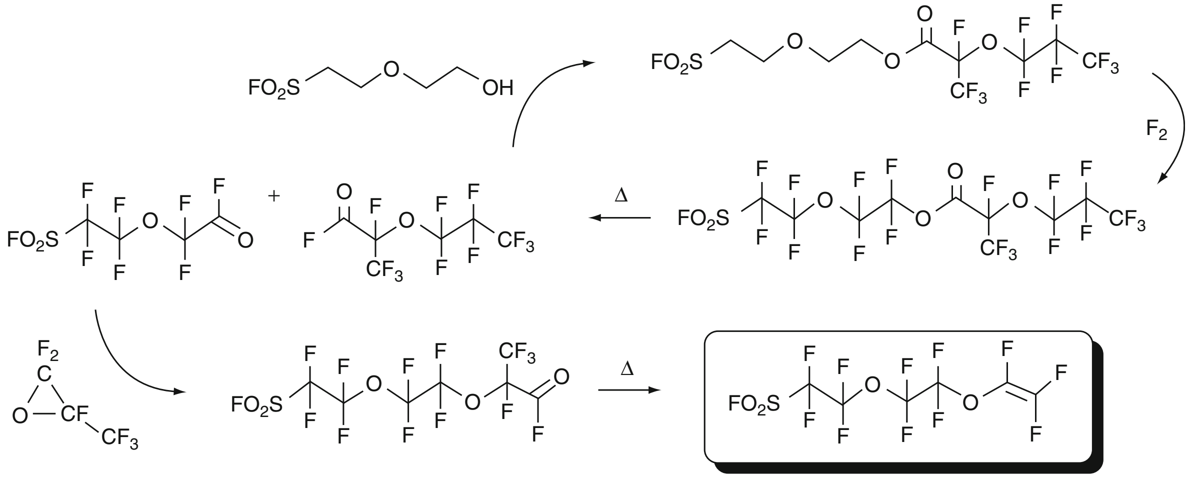
$$
\begin{aligned}
&Ce^{4+} + HO\cdot + H^+ \longrightarrow Ce^{3+} + H_2O\\
&Ce^{4+} + H_2O_2 \longrightarrow Ce^{3+} + HOO\cdot + H^+\\
&Ce^{4+} + HOO\cdot \longrightarrow Ce^{3+} + O_2 + H^+\\
&Ce^{3+} + 1/2\,H_2 \longrightarrow Ce^{4+} + H^+
\end{aligned}
$$

Membrane Electrolytes, from Perfluoro Sulfonic Acid (PFSA) to Hydrocarbon Ionomers. Figure 7
Radical quenching mechanism with cerium ions

Hydrocarbon Ionomer Membranes

High IEC Ionomers with Rigid Rod Backbone

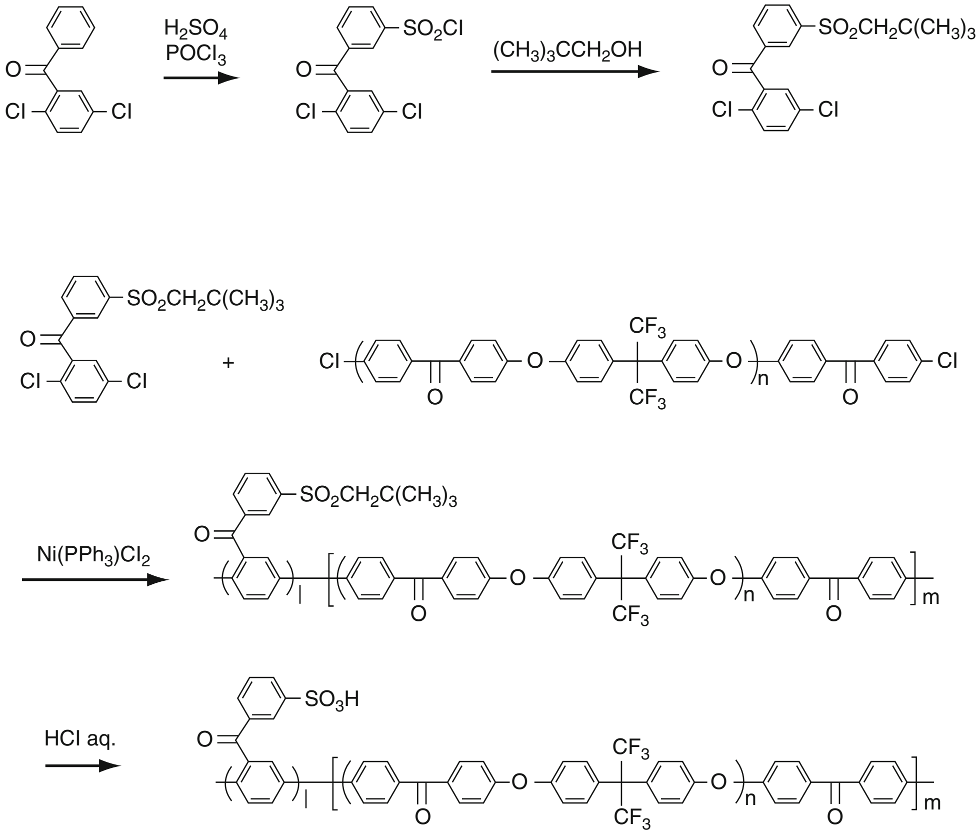
Poly(phenylene)s are probably one of the most attractive polymer backbones for hydrocarbon ionomers in terms of chemical stability and long-term durability since their main chains are composed of pure $C_{aromatic}$–$C_{aromatic}$ bonds, which afford rigid rod structure to the polymers. Most other hydrocarbon ionomers suffer from chemical degradation, which



Membrane Electrolytes, from Perfluoro Sulfonic Acid (PFSA) to Hydrocarbon Ionomers. Figure 8
Synthetic route of sulfonated perfluorovinyl monomers via direct fluorination of partially fluorinated compounds

often takes place at hetero linkages, especially electron-donating ether and aliphatic groups, under fuel cell operating conditions. The challenge is how to synthesize soluble high molecular weight poly(phenylene) ionomers (high enough to provide self-standing membranes) and how to introduce ionic groups. There are not many synthetic options available for poly(phenylene) derivatives compared to the other aromatic polymers with heteroatom linkages. Therefore, there have been limited number of reports for poly(phenylene) ionomers. Among them, poly (p-phenylene) ionomers developed by Goto et al. of JSR Corporation, Japan are one of the most successful examples. They have discovered that poly (p-phenylene)s with 3-sulfobenzoyl groups give appropriate properties as fuel cell membranes [23]. A typical synthetic approach is summarized in Fig. 9. The key monomer, neopentyl 3-(2,5-dichlorobenzoyl) benzenesulfonate, synthesized form 2,5-dichlorobenzophenone, was copolymerized with hydrophobic dichloro compounds via nickel-catalyzed coupling reaction. The neopentyl ester groups in the resulting polymers were removed via hydrolysis to obtain the title ionomers.

The striking feature of the poly(p-phenylene)-based ionomers is that the membranes obtained therefrom show well-developed hydrophilic/hydrophobic microphase separation. Such morphology can be controlled by (1) copolymer composition, (2) chemical structure of hydrophobic component, (3) sequenced structure and length of hydrophilic and hydrophobic components, and (4) membrane preparation conditions. Another characteristic of the poly (p-phenylene)-based ionomers is that the IEC can be higher than 3 meq/g without sacrificing good mechanical and chemical stability of the membranes. In Table 3 are summarized properties of typical JSR membranes. According to the disclosed data, the JSR membranes



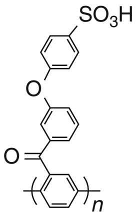
Membrane Electrolytes, from Perfluoro Sulfonic Acid (PFSA) to Hydrocarbon Ionomers. Figure 9
Synthesis of sulfonated poly(p-phenylene) copolymers

Membrane Electrolytes, from Perfluoro Sulfonic Acid (PFSA) to Hydrocarbon Ionomers. Table 3 Properties of the sulfonated poly(p-phenylene) copolymer (JSR) membranes

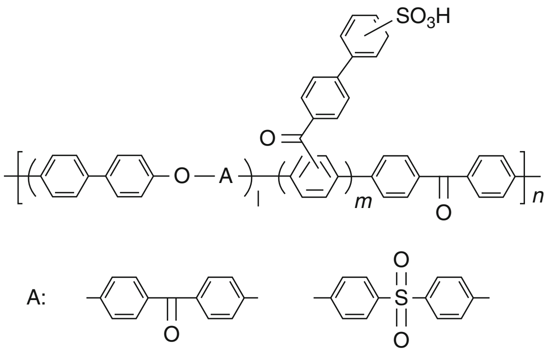
Proton conductivity	70°C, 80% RH	0.1 S/cm
	95°C, 80% RH	0.16 S/cm
Mechanical strength	Elongation at break at 23°C, 50% RH	100%
	Stress at break at 23°C, 50% RH	130 MPa
Gas permeability	H_2 at 80°C, dry	9 barrer
	O_2 at 40°C, 90% RH	6 barrer
Stability	IEC change after 1,000 h in 95°C water	0%
	Weight change after 1,000 h in 95°C water	0%

absorb more water, show higher proton conductivity, better thermal and mechanical stability, comparable hydrolytic stability, and much lower gas permeability compared to those of the conventional perfluorinated ionomer membranes. It has been recently claimed that the introduction of basic groups such as pyridine and imidazole as a third comonomer component could improve the durability of the poly(p-phenylene)-based ionomers. The JSR membranes have been successfully installed on Honda FCX Clarity fuel cell electric vehicles.

Rikukawa et al. of Sophia University have also developed a series of poly(p-phenylene) ionomers [24–26]. The main differences from the JSR membranes are that sulfophenylene groups are connected with ether bonds in the side chain because of the synthetic reason (Fig. 10). (According to JSR, such extra phenoxy groups do not have positive effect on the properties of the resulting ionomer membranes in terms of the proton conductivity and stability.) Rikukawa's group has applied post-sulfonation method so that the phenyl groups need to be activated with electron-donating ether groups. The advantage is that synthesis of monomers and high molecular weight polymers is easier. Their membranes showed similar properties to the JSR membranes, supporting the validity of the strategy to utilize rigid rod-like main chains.

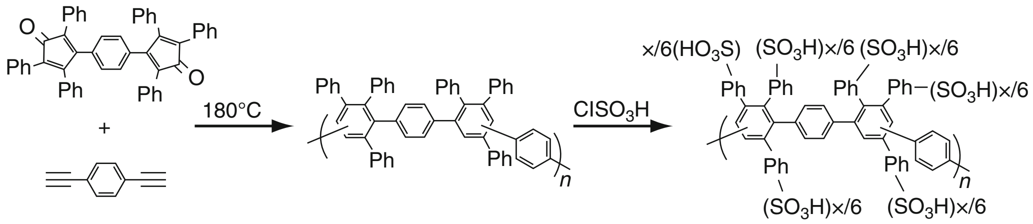


Membrane Electrolytes, from Perfluoro Sulfonic Acid (PFSA) to Hydrocarbon Ionomers. Figure 10 Sulfonated poly(p-phenylene) containing phenoxy side chains

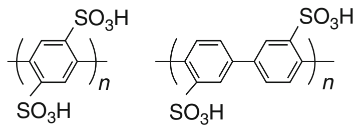


Membrane Electrolytes, from Perfluoro Sulfonic Acid (PFSA) to Hydrocarbon Ionomers. Figure 11 Multiblock copolymer of sulfonated poly(p-phenylene) and poly(arylene ether sulfone)

Sulfonated poly(2,5-benzophenone)s, derivatives of poly(p-phenylene)s with sulfobiphenylenecarbonyl side chains, were synthesized by McGrath and Ghassemi of Virginia Polytechnic Institute and State University [27]. The base (unsulfonated) polymers were thermally stable up to ca. 480°C in air and nitrogen, however, their film-forming capability was insufficient due to rather low molecular weight. They have also synthesized multiblock copolymers composed of the sulfonated oligo(2,5-benzophenone)s and oligo (arylene ether sulfone)s to achieve higher molecular weight and better film-forming capability (Fig. 11) [28]. The block copolymers showed high glass transition temperature (225°C) due to the arylene ether sulfone units. Proton conductivity of the membrane (IEC = 1.20 meq/g) was measured under specific conditions (in water at 30°C) to be 0.036 S/cm.



Membrane Electrolytes, from Perfluoro Sulfonic Acid (PFSA) to Hydrocarbon Ionomers. Figure 12
Synthesis of polysulfophenylated poly(phenylene) via Diels–Alder polymerization followed by sulfonation



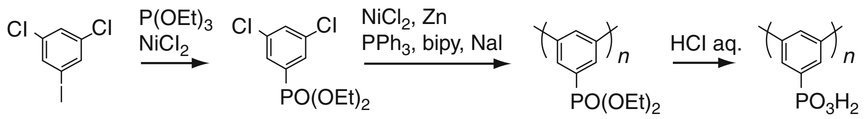
Membrane Electrolytes, from Perfluoro Sulfonic Acid (PFSA) to Hydrocarbon Ionomers. Figure 13
Poly(p-phenylene)s with sulfonic acid groups on the main chains

Cornelius et al. have synthesized a series of unique poly(phenylene)-based polyelectrolytes by Diels–Alder polymerization followed by post-sulfonation (Fig. 12) [29–32]. The ionomers are composed of sulfonated, highly phenylated poly(phenylene)s and do not carry any heteroatoms as their constituents except for the sulfonic acid groups. The complete aryl backbone resulted in a tough rigid-rod material with no Tg below the decomposition temperature. The stiffness of the ionomer backbone did not negatively affect the membrane properties such as water uptake (21–137%, in water) and proton conductivity (13–123 mS/cm, in water at 30°C) with IECs ranging from 0.98 to 2.2 meq/g.

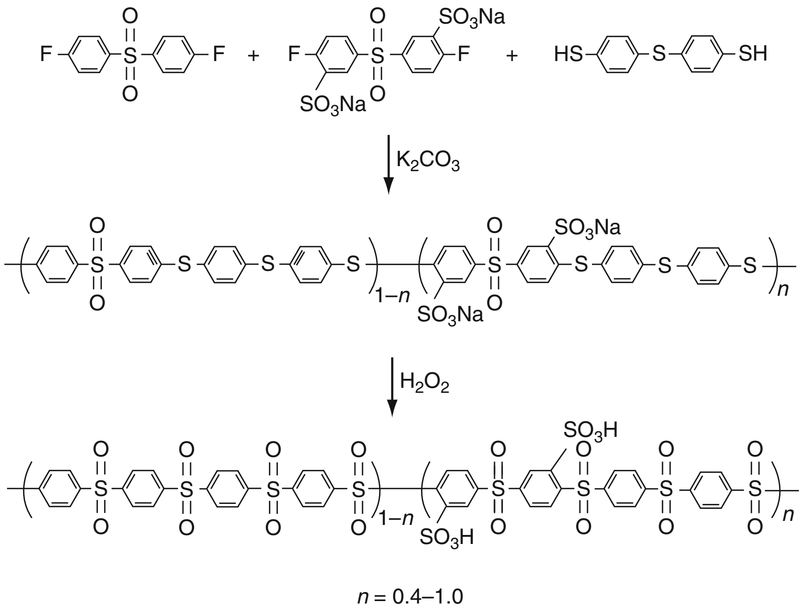
More recently, poly(p-phenylene) ionomers with simpler chemical structure has been proposed by the group of Litt of Case Western Reserve University (Fig. 13). They have synthesized poly(p-phenylene)s with sulfonic acid groups directly attached on the main chains via Ullmann coupling reaction. Their membranes seem promising for high-temperature operable fuel cells in terms of the proton conductivity; the proton conductivity of the membranes was 0.1 S/cm at 75°C and 15% RH, which was approximately 3 orders of magnitude higher than that of Nafion membrane and meets the requirements of DOE for the year of 2015. However, it is difficult to obtain high molecular weight polymers via the Ullmann coupling reaction, which causes poor mechanical properties of the membranes. This issue may be improved by copolymerizing with appropriate hydrophobic comonomers that are more reactive in the Ullmann coupling reaction and also give flexibility to the membranes.

Phosphonic acid containing poly(phenyelne)s were investigated by Kreuer's group of Max-Planck-Institute (Fig. 14) [33, 34]. m-Dichlorobenzene containing phosphonic acid ester was polymerized by nickel-catalyzed polycondensation reaction. The ester groups were hydrolyzed with acid to provide poly (m-phenylene phosphonic acid). Compared to the sulfonated poly(phenylene) ionomers, phosphonated analogues showed lower proton conductivity and better thermal stability due to the lower water affinity. The ionomer membranes are probably not suitable for fuel cells with dry or slightly humidified hydrogen as a fuel but may find applications using liquid fuels such as direct methanol fuel cells. In any case, the ionomers have to be cross-linked or incorporated with hydrophobic moieties in order to prevent excess swelling or dissolving in water.

A new synthetic approach has been developed by the same group to produce poly(phenylene) ionomers containing merely sulfone units connecting the phenylene rings ($SPSO_2$) (Fig. 15) [35, 36]. In the ionomers, each phenylene ring contains one sulfonic acid group (100% degree of sulfonation), which corresponds to very high IEC of 4.5 meq/g. The synthesis was carried out in two-step process, in which bis (sulfophenyl)sulfone was polymerized with sodium sulfide and then the subsequent oxidation reaction of sulfide linkages to sulfone gave the title ionomers. Since phenylene rings in the main chains are connected solely



Membrane Electrolytes, from Perfluoro Sulfonic Acid (PFSA) to Hydrocarbon Ionomers. Figure 14
Synthesis of poly(m-phenylene phosphonic acid)

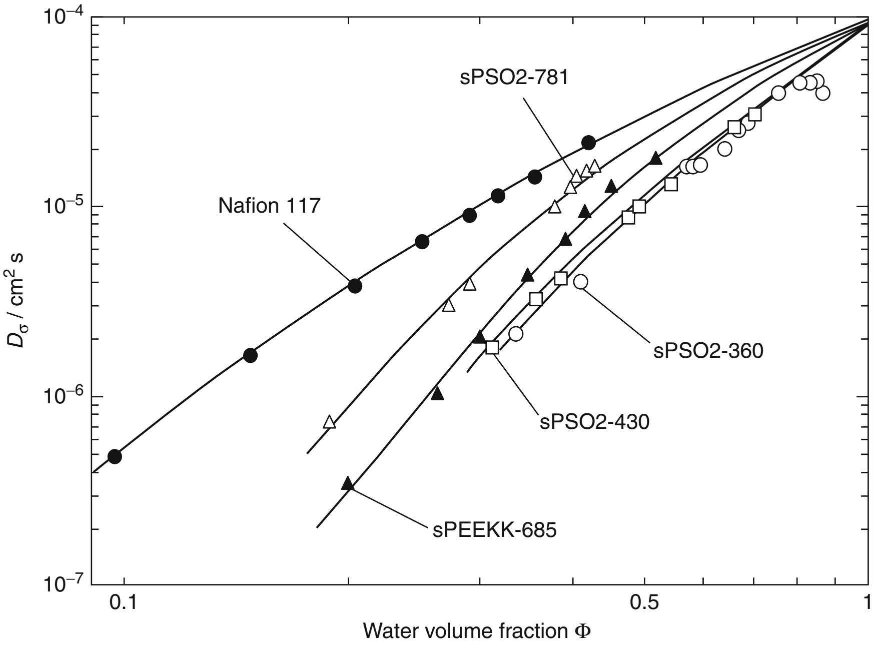


Membrane Electrolytes, from Perfluoro Sulfonic Acid (PFSA) to Hydrocarbon Ionomers. Figure 15
Synthesis of sulfonated poly(phenylene sulfone)s (sPSO2)

with strong electron-withdrawing sulfone groups and not with electron-donating groups such as ether and sulfide, the ionomers show high hydrolytic (no practical desulfonation at 180°C and high water activity) and thermooxidative stability (decomposition in air above 300°C). The high IEC afforded the ionomer membranes very high proton conductivity in the wide temperature range from 110°C to 160°C at a constant water vapor pressure (1 atm), which corresponds to ca. 50% and 15% RH, respectively. The proton diffusion coefficient (calculated from the conductivity data by use of Nernst–Einstein equation) increased with IEC values (Fig. 16). While the conductivity overcomes that of Nafion by a factor of 5–7, the ionomers are soluble in water and become brittle in the dry state. Approaches such as blends, graft-, and block-copolymers are under investigation. The materials have been registered as fumapem S (granular resin, solution, or dispersion) and fumion S (membranes) by fumatech company for commercialization.

Block Copolymers

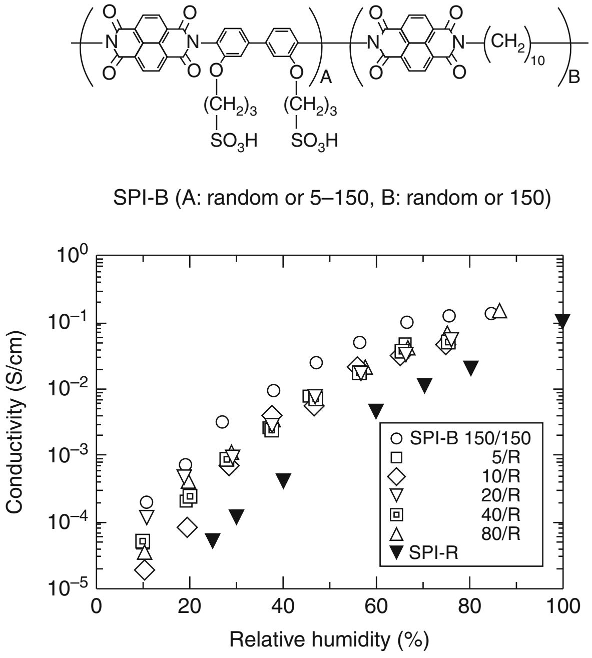
Block copolymers have been utilized in order for the hydrocarbon ionomer membranes to have well-developed and interconnected ionic domains. The strategy seems to work in many ionomers of different molecular structure. The block copolymers membranes show considerably higher proton conductivity than that of the random copolymer membranes with similar values of IEC. For example, Kawakami et al. of Tokyo Metropolitan University have proved that the block



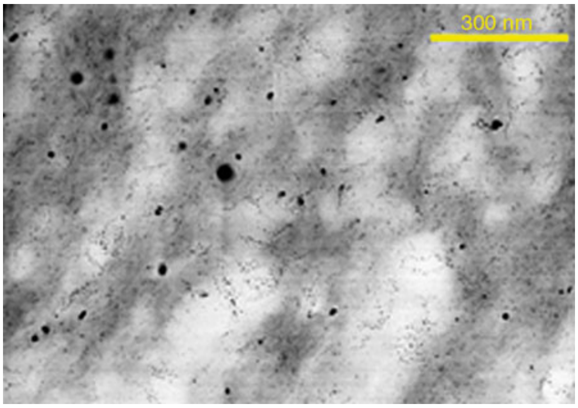
Membrane Electrolytes, from Perfluoro Sulfonic Acid (PFSA) to Hydrocarbon Ionomers. Figure 16
Proton diffusion coefficient of sPSO2, sulfonated poly(ether ether ketone ketone), and Nafion 117 membranes as a function of water volume fraction at 30°C (Reprinted from [36] with permission from the American Chemical Society)

copolyimides containing hexafluoropropylene groups in the hydrophobic segment are much more proton conductive than the random equivalents [37–40]. The differences in the conductivity were much more pronounced at low humidity. The proton conductivity of the block copolyimide membranes depended strongly upon the block chain lengths. The longer the block length was, the higher the conductivity became. They ascribe this effect to ionic channels of which formation depends on the block chain length.

Similar effect was confirmed by the group of Miyatake et al. of University of Yamanashi with other series of sulfonated polyimides of different main chain structure [41]. They have investigated the effect of block copolymer architecture on sulfonated polyimides containing aliphatic segments in the hydrophobic main chains and in the hydrophilic side chains. The block copolymer with longest block segments (the number of repeating unit was 150 for both hydrophilic and hydrophobic blocks) showed the highest proton conductivity of 2 × 10^{-2} S/cm at 80°C and 48% RH, which was comparable to that of the conventional perfluorinated ionomer membrane (Fig. 17). Well-connected



Membrane Electrolytes, from Perfluoro Sulfonic Acid (PFSA) to Hydrocarbon Ionomers. Figure 17
Humidity dependence of the proton conductivity of sulfonated block copolyimide membranes at 80°C



Membrane Electrolytes, from Perfluoro Sulfonic Acid (PFSA) to Hydrocarbon Ionomers. Figure 18
Scanning transmission electron microscopic image of SPI-B 150/150

hydrophilic domains, which looked a cloudlike belt with several hundred nanometers in width, were observed in the transmission electron microscopic images (Fig. 18).

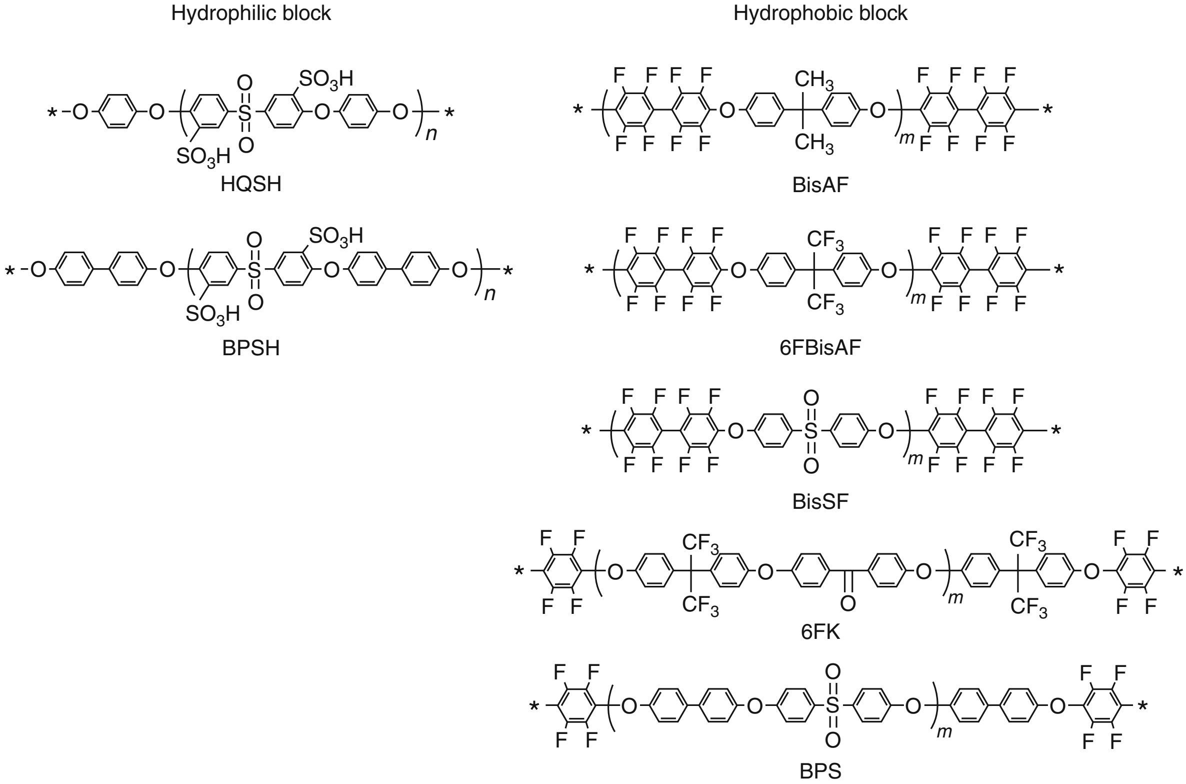
McGrath's group has done an extensive and systematic study on their sulfonated poly(arylene ether sulfone) block copolymers (Fig. 19) [42–51]. The block copolymers are composed of biphenol-based disulfonated arylene ether sulfone (so-called BPSH) units and the unsulfonated equivalents (BPS). The investigated properties of the block copolymer membranes include synthetic details with different main chain linkages, spectroscopic analyses of the chemical structure, water uptake, diffusion of water, proton conductivity at wide range of temperature and humidity, thermal transition and decomposition, morphological analyses (atomic force microscopic (AFM) and transmission electron microscopic (TEM) imaging, and small-angle X-ray scattering (SAXS) profile), mechanical strength, and fuel cell performance. A brief summary is

- High molecular weight block copolymers were obtained when reactive perfluorinated linkage (nonafluorobiphenylene or pentaflurophenylene) groups were attached at the both ends on unsulfonated hydrophobic oligomers.
- The linkage groups had some effect on the membrane properties. The fluorinated biphenylene groups seemed to promote nanophase separation, and thus water uptake and proton conductivity at low humidity than the fluorinated phenylene groups.
- The block copolymer membranes performed much better as proton exchange membranes for fuel cells than the random copolymers with similar IEC, especially in terms of proton conductivity at low humidity (on the order of mS/cm at 80°C and 30% RH, IEC = 1.5 meq/g).
- The block length rather than the IEC was more important to dominate water uptake and proton conductivity, where longer block length led to higher water uptake and higher proton conductivity. So were the nanophase separation (or connection of hydrophilic domains) (Fig. 20) and water diffusion coefficient.
- In the hydrated block copolymers, more freezing water (free and loosely bound water to sulfonic acid groups) existed than in the random copolymers due to the developed morphology (Table 4). The block length should be longer than 10 kDa in order to have noticeable improvement on the morphological order and proton conductivity.
- In addition to the block length and IEC values, casting conditions (such as the solvent and drying temperature or solvent removal rate) did have significant impact on membrane morphology and properties.
- The block copolymer membrane performed in an H_2/air fuel cell at 100°C and 40% RH comparable to Nafion membrane.

These findings are, more or less, applicable to the other hydrocarbon ionomers and useful to tailor the higher order structure and properties of ionomer membranes.

Polymers with Sulfonic Acid Clusters

In order to overcome the trade-off relationship between water uptake and proton conductivity (high conductivity can be achieved with high water absorbing membranes), Hay's group at McGill University has proposed a unique strategy. They have synthesized poly(arylene ether)s containing nanoclusters of up to 18 sulfonic acid groups either tethered on the end groups or distributed in the main chains (Fig. 21) [52–54, 55–58]. Membranes therefrom showed

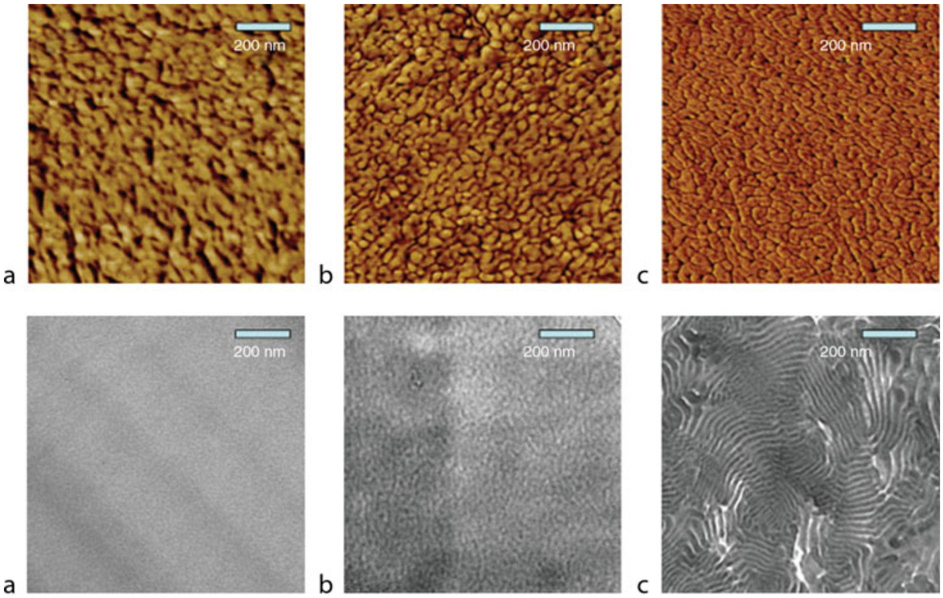


Membrane Electrolytes, from Perfluoro Sulfonic Acid (PFSA) to Hydrocarbon Ionomers. Figure 19
Chemical structure of BPSH block copolymers and their derivatives

significantly phase-separated (worm-like or large hydrophilic domains) and highly connected morphology in the TEM images. The highest proton conductivity at 120°C was 3.1 mS/cm at 10% RH and 10.5 mS/cm at 30% RH for a membrane with IEC = 1.0 meq/g. The reported conductivity values were extremely high compared to those of the other hydrocarbon ionomer membranes with such low IEC and even higher than those of Nafion under the same conditions, 1.7 mS/cm (at 10% RH) and 4.0 mS/cm (at 30% RH), respectively. The membrane absorbed ca. double amount of water compared to the Nafion membrane, which is thought to be responsible for the high proton conductivity.

Ueda and his coworkers of Tokyo Institute of Technology confirmed this strategy with their sulfonated poly(arylene ether sulfone)s (Fig. 22) [59–61]. Their polymers contain highly sulfonated moieties (up to ten sulfonic acid groups per repeating unit) randomly distributed in the main chains. Large difference in the polarity between highly sulfonated units and hydrophobic units caused the formation of defined phase-separated structures and well-connected proton pathways. The proton conductivity of the ionomer membrane with IEC = 2.38 meq/g was comparable to that of Nafion 117 at > 30% RH, 80°C.

Miyatake et al. have combined two strategies in a single polymer architecture (multiblock copolymers containing sulfonic acid clusters in their hydrophilic blocks) [62]. They have successfully synthesized a series of multiblock poly(arylene ether sulfone ketone)s (SPESKs) containing fully sulfonated fluorenylidene biphenylene units (Fig. 23) [63, 64]. The well-controlled post-sulfonation reaction of the precursor polymers enabled preferential sulfonation on each aromatic ring of the fluorenylidene biphenylene groups with 100% degree of sulfonation. The ionomer membranes showed unique morphology with well-developed hydrophilic/hydrophobic phase separation, depending on the block length of each segment. It was concluded that longer block length and/or higher IEC



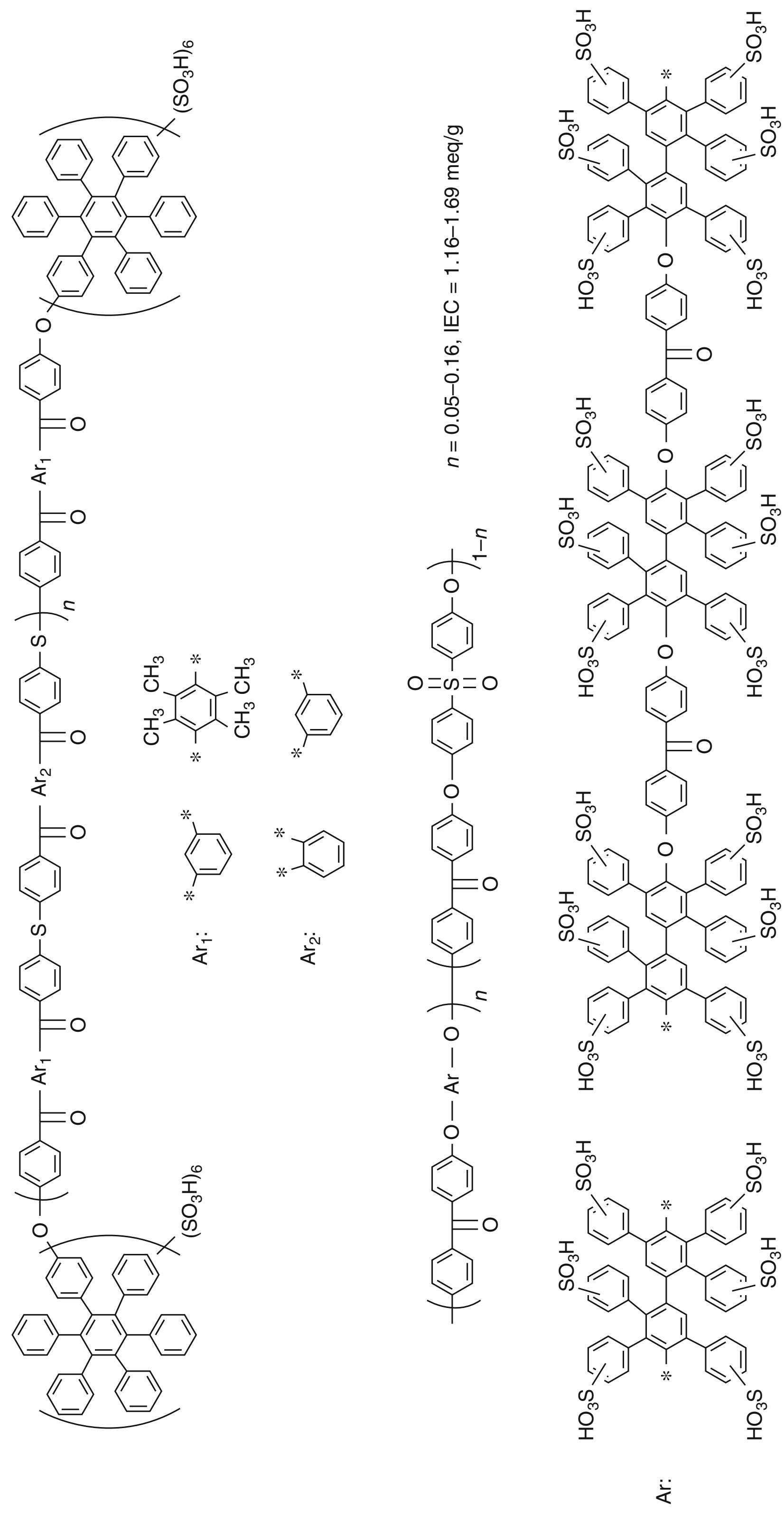
Membrane Electrolytes, from Perfluoro Sulfonic Acid (PFSA) to Hydrocarbon Ionomers. Figure 20
Atomic force microscopy (AFM) (top) and transmission electron microscopy (TEM) (bottom) images of BPSH-6FK multi-block copolymer membranes; (**a**) BPSH5-6FK5, (**b**) BPSH10-6FK10, and (**c**) BPSH10 for AFM or BPSH15 for TEM-6FK15. Numbers represent molecular weight of each block component in kDa (Reprinted from [47] with permission from Wiley Interscience)

Membrane Electrolytes, from Perfluoro Sulfonic Acid (PFSA) to Hydrocarbon Ionomers. Table 4 States of water molecules in the BPSH (random), BPSH-BPS (block), and Nafion membranes

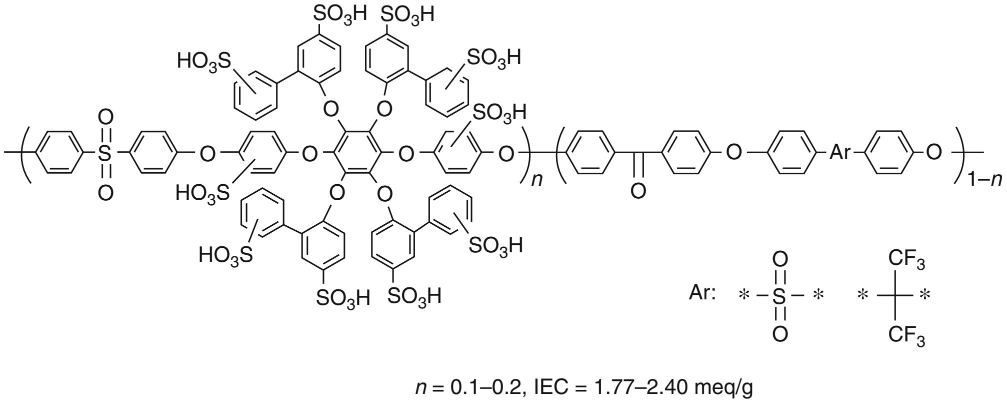
Membrane	IEC (meq/g)	Total water (λ^a)	Tightly bound water (λ^a)	Loosely bound water (λ^a)	Free water (λ^a)
BPSH30	1.34	12	4	8	0
BPSH3-BPS3	1.33	12	7	5	0
BPSH5-BPS5	1.39	13	8	5	0
BPSH10-BPS10	1.28	26	10	8	8
Nafion 112	0.90	17	3	9	5

[a]Number of water molecules absorbed per sulfonic acid group

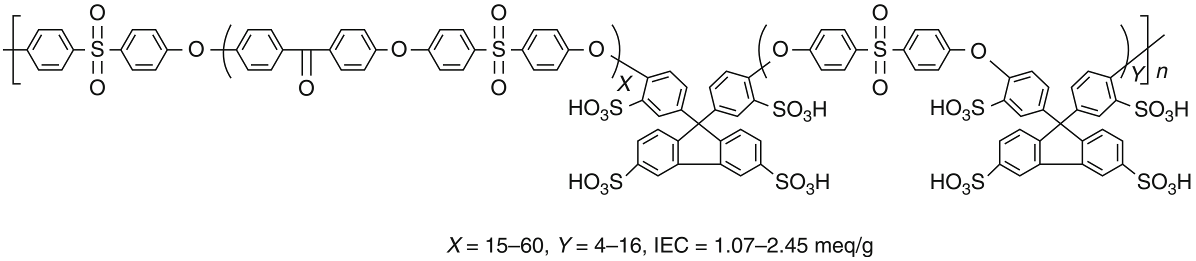
resulted in larger and better-connected hydrophilic clusters under dry conditions, while the morphology was less dependent on these factors under fully hydrated conditions. The multiblock copolymer membrane with IEC = 1.62 meq/g showed much higher proton conductivity than that of the random copolymer membrane with similar chemical structure and IEC (Fig. 24). The proton conductivity was similar or even higher compared to that of Nafion over a wide humidity range. The membrane retained high proton conductivity at 110°C. The high conductivity resulted from the high proton diffusion coefficient. Longer block length seemed effective in increasing proton diffusion coefficient, which coincided with the morphological observations. The multiblock copolymer membranes were stable to hydrolysis in water at



Membrane Electrolytes, from Perfluoro Sulfonic Acid (PFSA) to Hydrocarbon Ionomers. Figure 21
Poly(arylene ether)s containing sulfonic acid clusters



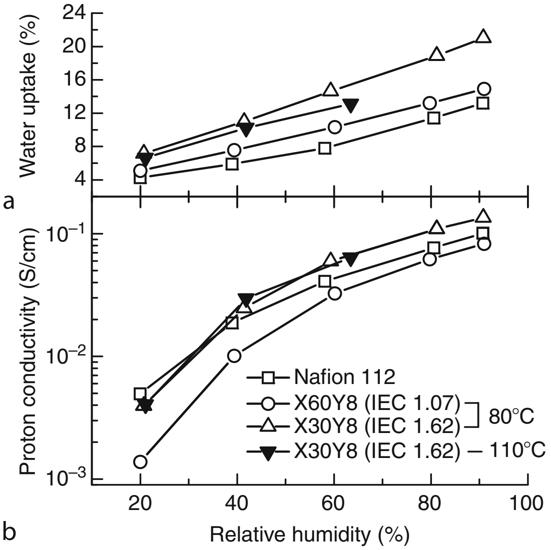
Membrane Electrolytes, from Perfluoro Sulfonic Acid (PFSA) to Hydrocarbon Ionomers. Figure 22
Poly(arylene ether sulfone)s with up to ten sulfonic acid groups per repeating unit



Membrane Electrolytes, from Perfluoro Sulfonic Acid (PFSA) to Hydrocarbon Ionomers. Figure 23
Multiblock poly(arylene ether sulfone ketone)s (SPESKs) containing fully sulfonated fluorenylidene biphenylene units

140°C for 24 h or at 100°C for 1000 h. The membranes degraded to some extent under harsh oxidative conditions (in Fenton's reagent), which is the fate of hydrocarbon ionomer compounds. Oxidative degradation is likely to occur at phenylene carbon atoms *ortho* to the ether bonds by the attack of highly oxidative hydroxyl radicals. The multiblock copolymer membrane showed much lower gas permeability than Nafion while humidity dependence of the permeability was different between hydrogen and oxygen. Hydrogen permeability showed reverse volcano-type dependence on the humidity, decreased slightly with humidification, and then increased with further humidification. Oxygen permeability simply increased with humidity. Such humidity dependency of oxygen permeability was similar to that of Nafion, however, not observed in the random copolymer membranes. The low gas permeability could mitigate their oxidative instability since hydrogen peroxide as a by-product is potentially less produced when the gas permeation through the membranes is low. A fuel cell was successfully operated with the multiblock copolymer membranes at 30% and 53% RH and 100°C (Fig. 25). The current density was 250 mA/cm^2 at 30% RH and 410 mA/cm^2 at 53% RH at a cell voltage of 0.6 V. The high proton conductivity of the membrane at low RH and high temperature was well confirmed in practical fuel-cell operation.

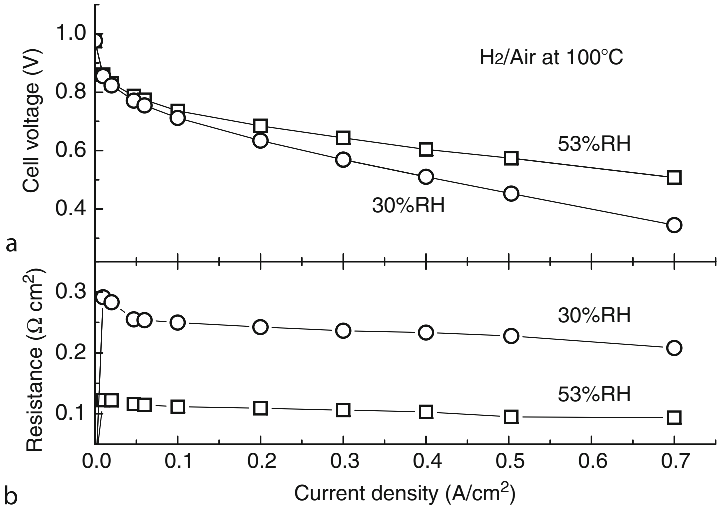
Zhang et al. of Changchun Institute of Applied Chemistry have also confirmed the similar strategy with their unique multiblock copolyimide ionomers, in which hydrophilic blocks were composed of disulfonated dianhydride and disulfonated diamine (each aromatic ring was sulfonated) (Fig. 26) [65]. While they have not done morphological investigations, the high concentration of the sulfonic acid groups in the hydrophilic block was effective in enhancing water absorbability and proton conductivity of the polyimide ionomer membranes (Fig. 27). For example, their high IEC (2.69 meq/g) block copolymer membranes were much more proton conductive than the random equivalent and the highest conductivity at 50% RH and 70°C was 3.2 × 10^{-2} S/cm for the multiblock copolyimide with the 50 repeating units in the hydrophilic blocks. The multiblock copolyimides showed lower proton concentration as a consequence of the higher water uptake, however, higher proton mobility compensated the proton concentration. Hydrolysis remains an issue for the polyimides and the block copolymer architecture did not help improve the hydrolytic stability (the fact is that the block copolymers were more susceptible to hydrolysis due to the higher water uptake).



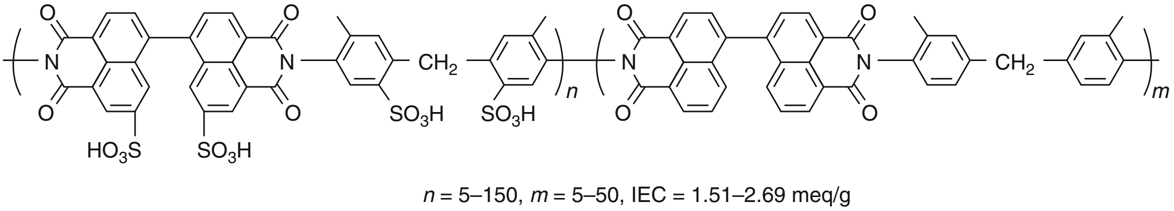
Membrane Electrolytes, from Perfluoro Sulfonic Acid (PFSA) to Hydrocarbon Ionomers. Figure 24
(**a**) Water uptake and (**b**) proton conductivity of Nafion 112 and block SPESKs (where IEC is in meq/g) as a function of relative humidity at 80 and 110°C (Reprinted from [64] with permission from Wiley Interscience)

Polymers with Pendant Acidic Groups

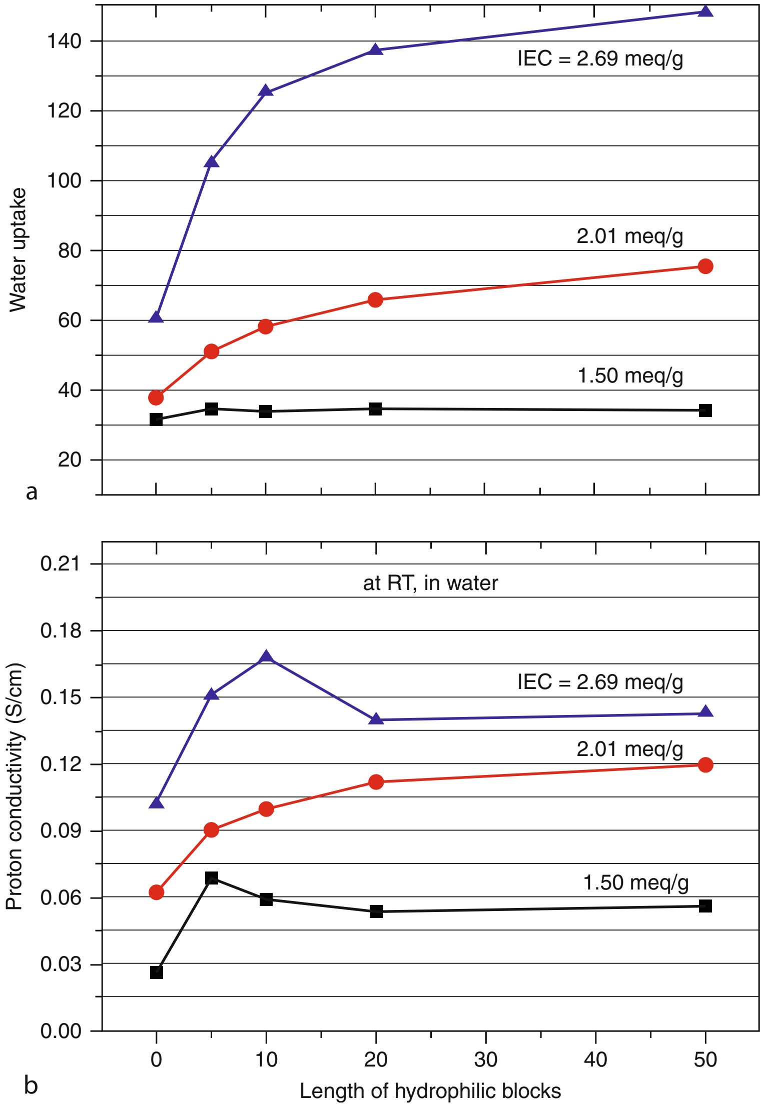
One of the reason for Nafion to have well-developed phase separation and interconnected ionic channels lies on its chemical structure that contains acid groups at the end of flexible perfluoroalkyl ether side chains. The introduction of pendant sulfonic acid groups has been investigated for the aromatic polymers such as polybenzimidazoles, polyimides, and poly(arylene ether)s, to confirm whether a similar effect can be obtained without perfluorinated chains (Fig. 28).



Membrane Electrolytes, from Perfluoro Sulfonic Acid (PFSA) to Hydrocarbon Ionomers. Figure 25
(**a**) H_2/air fuel cell performances including ohmic resistance voltage losses and (**b**) ohmic resistances of SPESK X30Y8 (IEC = 1.62 meq/g) at 100°C with humidification at 53 and 30% RH for both electrodes (Reprinted from [64] with permission from Wiley Interscience)



Membrane Electrolytes, from Perfluoro Sulfonic Acid (PFSA) to Hydrocarbon Ionomers. Figure 26
Multiblock copolyimides with highly sulfonated hydrophilic block



Membrane Electrolytes, from Perfluoro Sulfonic Acid (PFSA) to Hydrocarbon Ionomers. Figure 27
(**a**) Water uptake and (**b**) proton conductivity of multiblock copolyimides with different IEC as a function of hydrophilic block length (Reprinted from [65] with permission from Elsevier)

In most cases, flexible side chains between the aromatic main chains and the acid groups improve morphology and proton conductivity but not as striking as that in Nafion. For example, Jiang et al. of Jilin University have developed a series of poly (arylene ether sulfone)s containing long pendant acidic (sulfopropylbenzoyl, sulfopropoxybenzoyl, or sulfophenoxybenzoyl) groups [66–71] (a), in which membranes ionic clusters of ca. 10–20 nm in diameter were confirmed via TEM observation but rather isolated from one another. When relatively higher sulfonic acid concentration was applied, ionic clusters became larger (ca. 40 nm in diameter) and better interconnected. Poly(arylene ether sulfone)s containing pendant sulfonaphthyl groups (b) were synthesized by Manthiram's group at the University of Texas at Austin [72]. Their membranes outperformed Nafion 115 membrane in DMFC operation, while the effect of pendant naphthyl groups was not well investigated. Jannasch et al. of Lund University proposed poly(arylene ether sulfone)s tethered with hyper-acidified pendant groups (two or more sulfonic or phosphonic acid groups per pendant unit) (c) [73, 74]. By concentrating acidic groups locally onto the side chains, a distinct phase separation between hydrophobic polymer main chains and hydrophilic pendant groups was achieved. The water uptake of these kinds of ionomer membranes seemed rather high, and thus cross-linking or incorporation of inorganic additives may be needed to avoid excess swelling.

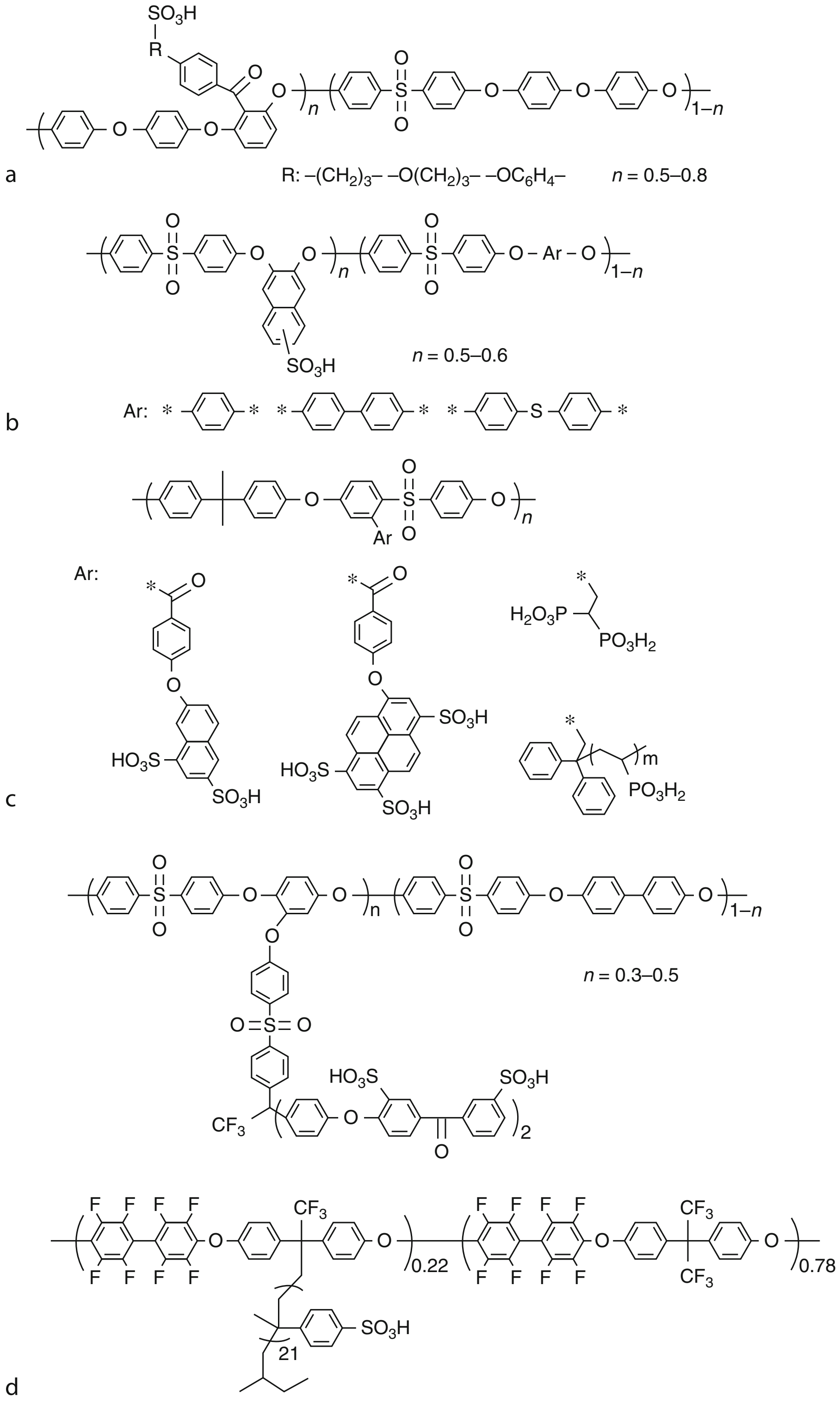
Recently, more developed ideas on the pendant acidic groups have been proposed by several research groups. Such include graft copolymers or comb-shaped ionomers, in which ionic grafts were substituted onto aromatic main chains. Jannasch et al. have further developed their above ideas by grafting poly(vinylphosphonic acid) (PVPA) onto poly(phenylene ether sulfone)s (c) [75]. Such molecular architecture caused phase separation and dual glass transition temperature due to the inherent miscibility of the stiff and hydrophobic polymer backbones and strongly hydrogen-bonded phosphonic acid side chains. The membrane with 57 wt% of the PVPA content showed 4.6 mS/cm (dry) and 93 mS/cm (100% RH) at 120°C. Still, the water uptake of the membrane was rather high and should be improved.

Guiver et al. of National Research Council, Canada developed comb-shaped poly(arylene ether) electrolytes containing 2–4 sulfonic acid groups on aromatic side chains (d) [76]. Their membranes showed relatively high proton conductivity and well-developed and continuous ionic domains. However, trade-off relationship between water uptake and proton conductivity of their membranes was not better than that of Nafion. In order to pronounce the hydrophilic/hydrophobic differences, another series of comb-shaped aromatic ionomers with highly fluorinated main chains and flexible poly(α-methyl styrene sulfonic acid) side chains were developed [77]. The membranes seemed to have better properties than their previous version, however, chemical instability of the side chains needed to be improved.

Hydrocarbon Polymers with Superacid Groups

One of the significant differences between hydrocarbon ionomers and perfluorosulfonic acid polymers is acid groups. The *pKa* value of benzenesulfonic acid ($PhSO_3H$) is −2.5 and that of trifluoromethanesulfonic acid (CF_3SO_3H) is −13. The pK_a value was estimated to be ca. −1 for sulfonated polyether ketone and ca. −6 for Nafion membranes [78]. Therefore, the effective proton concentration and proton mobility should be lower in the hydrocarbon ionomer membranes. Without appropriate molecular design such as multiblock copolymer and sulfonic acid clusters as mentioned above, hydrocarbon ionomer membranes lack well-developed ionic channels due to less pronounced hydrophilic and hydrophobic phase separation, which causes the lower proton conductivity at low humidity.

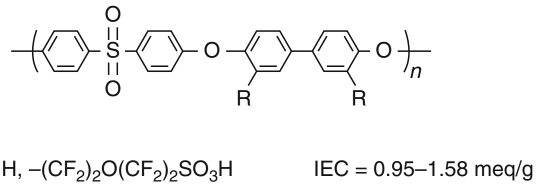
Yoshimura and Iwasaki of Sumitomo Chemical Co. have synthesized aromatic ionomers containing pendant perfluorosulfonic acid groups (Fig. 29) [79]. Poly (arylene ether sulfone) was brominated and then perfluorosulfonated via Ullmann coupling reaction in the presence of copper catalyst. The IEC was controllable up to 1.58 meq/g. The obtained ionomer membranes behaved very differently from the typical sulfonated aromatic ionomer membranes. Characteristic hydrophobic/hydrophilic separation (ca. 3–4 nm) was observed in the small-angle X-ray scattering (SAXS)



Membrane Electrolytes, from Perfluoro Sulfonic Acid (PFSA) to Hydrocarbon Ionomers. Figure 28
Aromatic polymers with pendant acidic groups

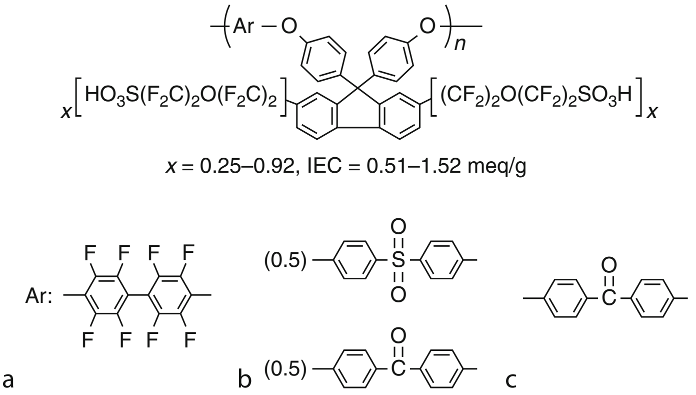
analyses of the hydrated samples. The ionic clusters were slightly smaller than those of Nafion membrane. The superacid groups were effective in improving thermal and mechanical properties of poly(arylene ether sulfone) ionomers as confirmed by dynamic mechanical analyses (DMA) and tensile testings. The superacidified ionomer membranes showed several times higher proton conductivity than that of the typical sulfonated poly(arylene ether sulfone) ionomers with similar IEC at 80°C, 50–90% RH.

Miyatake et al. have also confirmed the positive effect of superacid groups on the properties of the poly(arylene ether) ionomers [80, 81]. They have synthesized poly(arylene ether)s containing pendant superacid groups on fluorenyl groups (FSPEa-c) (Fig. 30) via similar method as that of Sumitomo Chemical Co. The superacid-containing ionomer membranes showed similar thermal and gas permeation properties to those of the conventional sulfonated aromatic ionomers. Instead, their morphology was more similar to that of Nafion. Well-developed hydrophilic/hydrophobic phase separation was observed, while the hydrophilic clusters were somewhat smaller than those of Nafion. More significant difference was observed in water uptake and proton conductivity (Fig. 31). For example, proton conductivity of superacid-containing ionomer membrane (IEC = 1.40 meq/g) was ca. 2 mS/cm at 80°C and 20% RH, which was considerably higher than that (0.02 mS/cm) of conventional sulfonated poly(arylene ether)s (SPE, IEC = 1.59 meq/g) under the same conditions. The two membranes showed very similar water uptake behavior at a wide range of humidity. The results imply that the superacid groups utilize water molecules more efficiently for proton conduction than arylsulfonic acid groups. Their study revealed that the main chain structure seemed to affect the properties and could be optimized for further improvement of the properties. The membrane showed good fuel cell performance at 80°C and 78 or 100% RH, however, the performance became lower under lower humidity conditions.

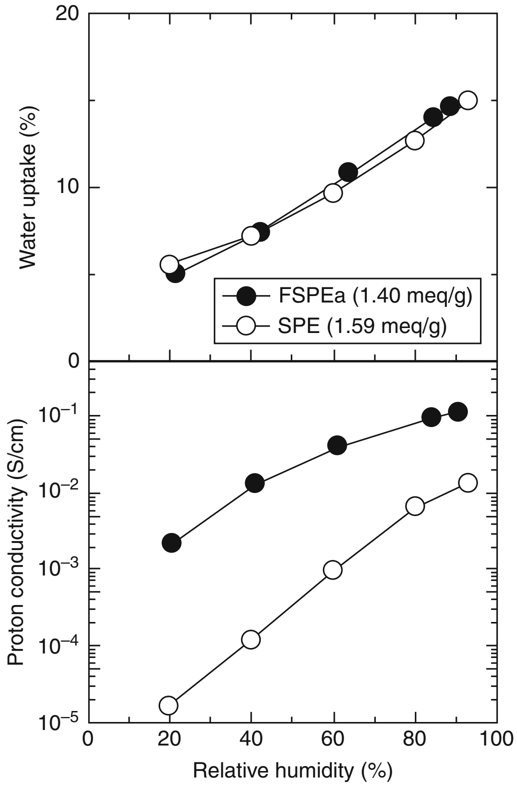


Membrane Electrolytes, from Perfluoro Sulfonic Acid (PFSA) to Hydrocarbon Ionomers. Figure 29
Poly(arylene ether sulfone)s with superacid groups

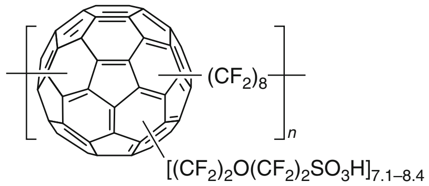
Hirakimoto et al. of Sony Co. have developed a unique ionomer containing superacid-substituted fullerene groups in the main chain (Fig. 32) [82]. Fullerene was substituted by averaging 7.1–8.4 superacid groups and were polymerized and cross-linked with 1,8-diiodoperfluorooctane to give the polymer with IEC = 1.88 meq/g. The ionomer was thermally stable up to 240°C as confirmed by TG analyses in dry nitrogen and showed high proton conductivity (2×10^{-2} S/cm) at 25°C and 50% RH as a pellet. Since the material did not seem to have good film-forming capability by



Membrane Electrolytes, from Perfluoro Sulfonic Acid (PFSA) to Hydrocarbon Ionomers. Figure 30
Poly(arylene ether)s containing superacid groups on fluorenyl groups



Membrane Electrolytes, from Perfluoro Sulfonic Acid (PFSA) to Hydrocarbon Ionomers. Figure 31 Humidity dependence of water uptake and proton conductivity of fluorene-containing poly(arylene ether)s with superacid groups (FSPEa) and conventional sulfonic acid groups (SPE) at 80°C



Membrane Electrolytes, from Perfluoro Sulfonic Acid (PFSA) to Hydrocarbon Ionomers. Figure 32 Superacid-modified polyfullerene

itself, composite membranes with poly(vinylidene difluoride) (PVDF) were investigated. The composite membrane containing 20 wt% of PVDF outperformed Nafion membrane in DMFC operation. The maximum power density of ca. 110 mW/cm^2 was achieved.

Composite Membranes

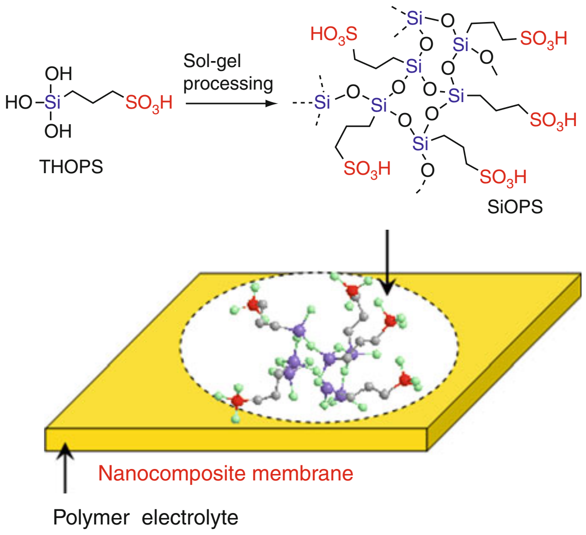
It has been proposed that modification of ionomer materials with hygroscopic oxides such silica, titania, tungsten oxide, zirconia, and zirconia phosphate could improve membrane properties. Introduction of nanoparticles of such oxides has been claimed to increase water affinity and reduce evaporation of water at high temperature and/or low humidity. The original research stems from Nafion composites [83–85] and recently, considerable effort has been devoted to apply the methodology to hydrocarbon ionomers. Some of the composite membranes showed better mechanical properties and lower gas or methanol permeability, however, the composite effect on the proton conductivity was not very obvious or even negative in some cases. In addition to the chemical and physical state of the inorganic additives, membrane fabrication procedure also seems a crucial factor since it affects membrane morphology and miscibility of inorganic additives with the matrix. Therefore, detailed and careful analyses are required to discuss the effect of additives. It is beyond the scope of this chapter to review the composite membranes. A few of the recent examples are introduced.

The group of GKSS Research Center Germany extensively researched the effect of a variety of inorganic nanoparticles on the properties of aromatic ionomer membranes [86–93]. Composite membranes were prepared from silicates and sulfonated poly(ether ketone)s or sulfonated poly(ether ether ketone)s. For DMFC applications, the composite membranes showed promising properties with lower methanol and water permeability and comparable (or higher) proton conductivity compared to the parent polymer membranes. The flux of water and methanol decreased with the increase in content of silicates.

In the case of sulfonated poly(phthalazinone ether ketone)s being used as a matrix, composite membranes showed less swelling in water and methanol, better mechanical and thermal stability, but lower proton conductivity [94, 95]. The results suggest that the amount of inorganic additives have to be optimized for each ionomer material and fuel cell operation conditions. Addition of 5 wt% of silica was optimum for poly(phthalazinone ether ketone)s to obtain high OCV and high power density of DMFC with 3 M methanol at 70°C.

It has been proposed that hybrid membranes are more proton conductive at low humidity than the parent polymer membranes. For example, zirconia phosphate (ZrP) was formed in the sulfonated poly (arylene ether sulfone). Nanoparticles of zirconia phosphate were homogeneously distributed in the membranes as crystalline α-zirconium hydrogen phosphate hydrate [96]. The composite membrane (with ZrP up to 50 wt%) showed 3.7 × 10^{-3} S/cm of the proton conductivity at 90°C, 30% RH, which was ca. 5 times higher than that of the parent polymer membrane under the same conditions. It is claimed that the composite effect depends upon interfacial contact between the polymer matrix and additives, and inappropriate preparation procedure could result in opposite effects.

In order to take more advantage of the effect of incorporated inorganic additives and to produce highly proton conductive composite membranes, Miyatake et al. have utilized 3-trihydroxysilyl-1-propanesulfonic acid (THOPS) as a precursor of the inorganic component (Fig. 33) [97]. THOPS has a sulfonic acid group covalently bonded to a silicon atom through aliphatic chain. The in situ sol–gel process in the polymer electrolyte matrices gave composite membranes containing sulfopropylated polysilsesquioxane ($\{SiO_{3/2}(CH_2)_3SO_3H\}_n$; SiOPS). Two different ionomer, sulfonated polyimides and sulfonated poly (arylene ether)s, were used as matrices. In the STEM images of the Ag-stained membrane samples, the ionic domains of approximately 5 nm in diameter were observed which were well-connected to each other. The connectivity of the ionic domains was significantly improved compared to that of the parent matrix membrane, while the size of the spherical ionic clusters did not alter. It is thought that the ionic domains result from the aggregation of sulfonic acid groups both from the matrix polymer and SiOPS. EDX analyses of the composite samples support highly dispersed SiOPS. These results suggest that well-dispersed nanocomposite membranes were formed. Nanocomposite membranes thus obtained showed much higher proton conductivity (up to 30 times) than that of the original membranes and accordingly less dependence of the conductivity upon the humidity. The ionomer properties such as thermal, hydrolytic and oxidative stability, and gas permeability were rather unaffected by the SiOPS. The methodology seems to be versatile as confirmed by the two different series of polymer electrolytes, although the miscibility with SiOPS depended on the matrix polymer.



Membrane Electrolytes, from Perfluoro Sulfonic Acid (PFSA) to Hydrocarbon Ionomers. Figure 33 Schematic representation of the proton-conductive nanocomposite membranes composed of ionomer (matrix) and sulfopropylated polysilsesquioxane (SiOPS) (Reprinted from [97] with permission from Wiley Interscience)

Future Directions

The success of fuel cell industry depends greatly on the membrane technology. Since the biggest market of fuel cells will lie in the electric vehicles, most of the effort on the development of proton exchange membranes has been and will be devoted to wide temperature and low humidity (or nominally dry) operable membranes. In the last decade, PFSAs have achieved a significant progress and probably meet most of the requirements for the commercialization of fuel cell vehicles. At least, reasonable performance and durability can be obtained under well-controlled operating conditions using relatively large balance of plant (BOP). On a long-term basis, the current situation is not acceptable and more effort has to be made in developing better (possibly, non-fluorinated) ionomer membranes that fulfill all the above-mentioned requirements. This indicates that investment in such research is essential.

There is no doubt that the PFSA membranes take initiative in this field and contribute a great deal in the commercialization and wide diffusion of fuel cells in the early stage. In terms of environmental compatibility (recyclability or disposability) and production cost, the PFSA membranes would better be replaced with non-fluorinated alternative materials within the next decade. Challenge is how to achieve comparable conductivity and durability with the non-fluorinated membranes. Unfortunately, no alternative materials have overcome the trade-off relationship between these two incompatible properties. Aromatic hydrocarbon ionomers are one of the possible candidates and run next to PFSAs in the race of membrane development. While a number of effective approaches have been proposed, there still remains a significant gap between them. Alternative membranes can show comparable conductivity to PFSAs by absorbing more water, however, such excessive hydrophilicity results in chemical and mechanical instability. All hydrocarbon membranes tend to show severer decrease of proton conductivity with decreasing RH compared to the conductivity of PFSA membranes. A technological breakthrough is certainly required.

Due to the limited space, the other important role of proton-conducting materials as a binder in the gas diffusion electrode has not been mentioned in this chapter. In order to maximize potential activity of catalysts (both for the oxygen reduction and hydrogen oxidation catalysts), the ionomer materials are expected to function differently from the membranes. In addition to the basic properties such as proton conductivity and durability, the binder materials are required to permeate reactant gases and discharge product water. Excess swelling under hydrated conditions often disturb oxygen from diffusing onto the catalyst surface. The membranes are generally flat and rather thick (>10 μm), while the binders cover uneven nanostructured catalysts on carbon black in several nm thickness. The ionomers may behave differently in such thin layers, however, there have been few research works on this topic, particularly for non-fluorinated ionomers. Although some non-fluorinated ionomers are claimed to perform comparably to Nafion as membranes, they show considerably lower performance as a binder in the gas diffusion electrode in most cases. These issues need to be addressed more carefully and extensively in order to realize high-performance fluorine-free fuel cells.

Bibliography

Primary Literature

1. http://www1.eere.energy.gov/hydrogenandfuelcells/mypp/
2. Arenz M, Schmidt TJ, Wandelt K, Ross PN, Markovic NM (2003) The oxygen reduction reaction on thin palladium films supported on a Pt(111) electrode. J Phys Chem B 107:9813–9819
3. Aieta NV, Stanis RJ, Horan JL, Yandrasits MA, Cookson DJ, Ingham B, Toney MF, Hamrock SJ, Herring AM (2009) Clipped random wave morphologies and the analysis of the SAXS of an ionomer formed by copolymerization of tetrafluoroethylene and CF2 = CFO(CF2)4SO3H. Macromolecules 42: 5774–5780
4. Emery M, Frey M, Guerra M, Haugen G, Hintzer K, Lochhaas KH, Pham P, Pierpont D, Schaberg M, Thaler A, Yandrasits M, Hamrock S (2007) The development of new membranes for proton exchange membrane fuel cells. ECS Trans 11:3–14
5. Arcella V, Troglia C, Ghielmi A (2005) Hyflon ion membranes for fuel cells. Ind Eng Chem Res 44:7646–7651
6. Kreuer KD, Schuster M, Obliers B, Diat O, Traub U, Fuchs A, Klock U, Paddison SJ, Maier J (2008) Short-side-chain proton conducting perfluorosulfonic acid ionomers: why they perform better in PEM fuel cells. J Power Sources 178:499–509
7. Yoshitake M, Watakabe A (2008) Perfluorinated ionic polymers for PEFCs (including supported PFSA). Adv Polym Sci, Fuel Cells I 215:127–155
8. Appleby AJ, Velev OA, LeHelloco JG, Parthasarthy A, Srinivasan S, DesMarteau DD, Gillette MS, Ghosh JK (1993) Polymeric perfluoro bis-sulfonimides as possible fuel cell electrolytes. J Electrochem Soc 140:109–111
9. Kotov SV, Pedersen SD, Qiu W, Qiu Z-M, Burton DJ (1997) Preparation of perfluorocarbon polymers containing phosphonic acid groups. J Fluorine Chem 82:13–19
10. Thomas BH, DesMarteau DD (2005) Self-emulsifying polymerization (SEP) of 3, 6-dioxa-Delta 7–4-trifluoromethyl perfluorooctyl trifluoromethyl sulfonimide with tetrafluoroethylene. J Fluorine Chem 126:1057–1064
11. Thomas BH, Shafer G, Ma JJ, Tu M-H, DesMarteau DD (2004) Synthesis of 3, 6-dioxa-Delta 7–4-trifluoromethyl perfluorooctyl trifluoromethyl sulfonimide: bis [(perfluoroalkyl)sulfonyl] superacid monomer and polymer. J Fluorine Chem 125:1231–1240
12. Uematsu N, Hoshi N, Koga T, Ikeda M (2006) Synthesis of novel perfluorosulfonamide monomers and their application. J Fluorine Chem 127:1087–1095
13. Coms FD (2008) The chemistry of fuel cell membrane chemical degradation. ECS Trans 16:235–255
14. Curtin DE, Lousenberg RD, Henry TJ, Tangeman PC, Tisack ME (2004) Advanced materials for improved PEMFC performance and life. J Power Sources 131:41–48

15. Schiraldi DA (2006) Perfluorinated polymer electrolyte membrane durability. Polym Rev 46:315–327
16. Danilczuk M, Perkowski AJ, Schlick S (2010) Ranking the stability of perfluorinated membranes used in fuel cells to attack by hydroxyl radicals and the effect of Ce(III): a competitive kinetics approach based on spin trapping ESR. Macromolecules 43:3352–3358
17. Endoh E (2008) Development of highly durable PFSA membrane and MEA for PEMFC under high temperature and low humidity conditions. ECS Trans 16:1229–1240
18. Ghassemzadeh L, Kreuer KD, Maier J, Muller K (2010) Chemical degradation of nation membranes under mimic fuel cell conditions as investigated by solid-state NMR spectroscopy. J Phys Chem C 114:14635–14645
19. Coms FD, Liu H, Owejan JE (2008) Mitigation of perfluorosulfonic acid membrane chemical degradation using cerium and manganese ions. ECS Trans 16:1735–1747
20. Danilczuk M, Schlick S, Coms FD (2009) Cerium(III) as a stabilizer of perfluorinated membranes used in fuel cells: in situ detection of early events in the ESR resonator. Macromolecules 42:8943–8949
21. Trogadas P, Parrondo J, Ramani V (2008) Degradation mitigation in polymer electrolyte membranes using cerium oxide as a regenerative free-radical scavenger. Electrochem Solid-State Lett 11:B113–B116
22. Okazoe T, Murotani E, Watanabe K, Itoh M, Shirakawa D, Kawahara K, Kaneko I, Tatematsu S (2004) An entirely new methodology for synthesizing perfluorinated compounds: synthesis of perfluoroalkanesulfonyl fluorides from nonfluorinated compounds. J Fluorine Chem 125:1695–1701
23. Goto K, Rozhanskii I, Yamakawa Y, Otsuki T, Naito Y (2008) Development of aromatic polymer electrolyte membrane with high conductivity and durability for fuel cell. Polym J 41:95–104
24. Bae JM, Honma I, Murata M, Yamamoto T, Rikukawa M, Ogata N (2002) Properties of selected sulfonated polymers as proton-conducting electrolytes for polymer electrolyte fuel cells. Solid State Ionics 147:189–194
25. Kobayashi T, Rikukawa M, Sanui K, Ogata N (1998) Proton-conducting polymers derived from poly(ether-etherketone) and poly(4-phenoxybenzoyl-1, 4-phenylene). Solid State Ionics 106:219–225
26. Yanagimachi S, Kaneko K, Takeoka Y, Rikukawa M (2003) Synthesis and evaluation of phosphonated poly (4-phenoxybenzoyl-1, 4-phenylene). Synth Met 135:69–70
27. Ghassemi H, McGrath JE (2004) Synthesis and properties of new sulfonated poly(p-phenylene) derivatives for proton exchange membranes. I. Polymer 45:5847–5854
28. Ghassemi H, Ndip G, McGrath JE (2004) New multiblock copolymers of sulfonated poly(4′-phenyl-2, 5-benzophenone) and poly(arylene ether sulfone) for proton exchange membranes. II. Polymer 45:5855–5862
29. Fujimoto CH, Hickner MA, Cornelius CJ, Loy DA (2005) Ionomeric poly(phenylene) prepared by Diels-Alder polymerization: synthesis and physical properties of a novel polyelectrolyte. Macromolecules 38:5010–5016
30. He L, Fujimoto CH, Cornelius CJ, Perahia D (2009) From solutions to membranes: structure studies of sulfonated polyphenylene ionomers. Macromolecules 42:7084–7090
31. Hickner MA, Fujimoto CH, Cornelius CJ (2006) Transport in sulfonated poly(phenylene)s: proton conductivity, permeability, and the state of water. Polymer 47:4238–4244
32. Stanis RJ, Yaklin MA, Cornelius CJ, Takatera T, Umemoto A, Ambrosini A, Fujimoto CH (2010) Evaluation of hydrogen and methanol fuel cell performance of sulfonated diels alder poly (phenylene) membranes. J Power Sources 195:104–110
33. Rager T, Schuster M, Steininger H, Kreuer K-D (2007) Poly(1, 3-phenylene-5-phosphonic acid), a fully aromatic polyelectrolyte with high ion exchange capacity. Adv Mater 19:3317–3321
34. Steininger H, Schuster M, Kreuer KD, Kaltbeitzel A, Bingoel B, Meyer WH, Schauff S, Brunklaus G, Maier J, Spiess HW (2007) Intermediate temperature proton conductors for PEM fuel cells based on phosphonic acid as protogenic group: a progress report. Phys Chem Chem Phys 9:1764–1773
35. Schuster M, de Araujo CC, Atanasov V, Andersen HT, Kreuer K-D, Maier J (2009) Highly sulfonated poly(phenylene sulfone): preparation and stability issues. Macromolecules 42:3129–3137
36. Schuster M, Kreuer K-D, Andersen HT, Maier J (2007) Sulfonated poly(phenylene sulfone) polymers as hydrolytically and thermooxidatively stable proton conducting ionomers. Macromolecules 40:598–607
37. Kashimura Y, Aoyama S, Kawakami H (2009) Gas transport properties of asymmetric block copolyimide membranes. Polym J 41:961–967
38. Nakano T, Nagaoka S, Kawakami H (2005) Preparation of novel sulfonated block copolyimides for proton conductivity membranes. Polym Adv Technol 16:753–757
39. Nakano T, Nagaoka S, Kawakami H (2006) Proton conductivity of sulfonated long-chain-block copolyimide films. Kobunshi Ronbunshu 63:200–204
40. Niwa M, Nagaoka S, Kawakami H (2006) Preparation of novel fluorinated block copolyimide membranes for gas separation. J Appl Polym Sci 100:2436–2442
41. Asano N, Miyatake K, Watanabe M (2006) Sulfonated block polyimide copolymers as a proton-conductive membrane. J Polym Sci, A: Polym Chem 44:2744–2748
42. Badami AS, Lane O, Lee H-S, Roy A, McGrath JE (2009) Fundamental investigations of the effect of the linkage group on the behavior of hydrophilic-hydrophobic poly(arylene ether sulfone) multiblock copolymers for proton exchange membrane fuel cells. J Membr Sci 333:1–11
43. Badami AS, Roy A, Lee H-S, Li Y, McGrath JE (2009) Morphological investigations of disulfonated poly(arylene ether sulfone)-b-naphthalene dianhydride-based polyimide multiblock copolymers as potential high temperature proton exchange membranes. J Membr Sci 328:156–164
44. Ghassemi H, McGrath JE, Zawodzinski TA (2006) Multiblock sulfonated-fluorinated poly(arylene ether)s for a proton exchange membrane fuel cell. Polymer 47:4132–4139

45. Lee H-S, Lane O, McGrath JE (2010) Development of multiblock copolymers with novel hydroquinone-based hydrophilic blocks for proton exchange membrane (PEM) applications. J Power Sources 195:1772–1778
46. Lee H-S, Roy A, Lane O, Dunn S, McGrath JE (2008) Hydrophilic-hydrophobic multiblock copolymers based on poly(arylene ether sulfone) via low-temperature coupling reactions for proton exchange membrane fuel cells. Polymer 49:715–723
47. Lee HS, Roy A, Lane O, Lee M, McGrath JE (2010) Synthesis and characterization of multiblock copolymers based on hydrophilic disulfonated poly(arylene ether sulfone) and hydrophobic partially fluorinated poly(arylene ether ketone) for fuel cell applications. J Polym Sci Pol Chem 48:214–222
48. Lee M, Park JK, Lee H-S, Lane O, Moore RB, McGrath JE, Baird DG (2009) Effects of block length and solution-casting conditions on the final morphology and properties of disulfonated poly(arylene ether sulfone) multiblock copolymer films for proton exchange membranes. Polymer 50:6129–6138
49. Roy A, Hickner MA, Einsla BR, Harrison WL, McGrath JE (2009) Synthesis and characterization of partially disulfonated hydroquinone-based poly(arylene ether sulfone)s random copolymers for application as proton exchange membranes. J Polym Sci Pol Chem 47:384–391
50. Roy A, Lee H-S, McGrath JE (2008) Hydrophilic-hydrophobic multiblock copolymers based on poly(arylene ether sulfone)s as novel proton exchange membranes - Part B. Polymer 49:5037–5044
51. Yu X, Roy A, Dunn S, Yang J, McGrath JE (2006) Synthesis and characterization of sulfonated-fluorinated, hydrophilic-hydrophobic multiblock copolymers for proton exchange membranes. Macromol Symp 245/246:439–449
52. Matsumura S, Hlil AR, Hay AS (2008) Synthesis, properties, and sulfonation of novel dendritic multiblock copoly(ether-sulfone). J Polym Sci Pol Chem 46:6365–6375
53. Matsumura S, Hlil AR, Lepiller C, Gaudet J, Guay D, Hay AS (2008) Ionomers for proton exchange membrane fuel cells with sulfonic acid groups on the end groups: novel linear aromatic poly(sulfide-ketone)s. Macromolecules 41:277–280
54. Matsumura S, Hlil AR, Lepiller C, Gaudet J, Guay D, Shi Z, Holdcroft S, Hay AS (2008) Ionomers for proton exchange membrane fuel cells with sulfonic acid groups on the end groups: novel branched poly(ether-ketone)s. Macromolecules 41:281–284
55. Matsumura S, Hlil AR, Al-Souz MAK, Gaudet J, Guay D, Hay AS (2009) Ionomers for proton exchange membrane fuel cells by sulfonation of novel dendritic multiblock copoly(ether-sulfone)s. J Polym Sci Pol Chem 47:5461–5473
56. Matsumura S, Hlil AR, Du N, Lepiller C, Gaudet J, Guay D, Shi Z, Holdcroft S, Hay AS (2008) Ionomers for proton exchange membrane fuel cells with sulfonic acid groups on the end-groups: novel branched poly(ether-ketone)s with 3, 6-ditrityl-9H-carbazole end-groups. J Polym Sci Pol Chem 46:3860–3868
57. Tian S, Meng Y, Hay AS (2009) Membranes from poly(aryl ether)-based ionomers containing multiblock segments of randomly distributed nanoclusters of 18 sulfonic acid groups. J Polym Sci Pol Chem 47:4762–4773
58. Tian S, Meng Y, Hay AS (2009) Membranes from poly(aryl ether)-based ionomers containing randomly distributed nanoclusters of 6 or 12 sulfonic acid groups. Macromolecules 42:1153–1160
59. Higashihara T, Matsumoto K, Ueda M (2009) Sulfonated aromatic hydrocarbon polymers as proton exchange membranes for fuel cells. Polymer 50:5341–5357
60. Matsumoto K, Higashihara T, Ueda M (2009) Locally and densely sulfonated poly(ether sulfone)s as proton exchange membrane. Macromolecules 42:1161–1166
61. Matsumoto K, Higashihara T, Ueda M (2009) Locally sulfonated poly(ether sulfone)s with highly sulfonated units as proton exchange membrane. J Polym Sci Pol Chem 47:3444–3453
62. Bae B, Miyatake K, Watanabe M (2009) Synthesis and Properties of Sulfonated Block Copolymers Having Fluorenyl Groups for Fuel-Cell Applications. ACS Appl Mater Interfaces 1:1279–1286
63. Bae B, Miyatake K, Watanabe M (2010) Sulfonated poly(arylene ether sulfone ketone) multiblock copolymers with highly sulfonated block. Synth Properties Macromol 43:2684–2691
64. Bae B, Yoda T, Miyatake K, Uchida H, Watanabe M (2010) Proton-conductive aromatic ionomers containing highly sulfonated blocks for high-temperature-operable fuel cells. Angew Chem Int Ed 49:317–320
65. Li N, Liu J, Cui Z, Zhang S, Xing W (2009) Novel hydrophilic-hydrophobic multiblock copolyimides as proton exchange membranes: enhancing the proton conductivity. Polymer 50:4505–4511
66. Liu B, Robertson GP, Kim D-S, Guiver MD, Hu W, Jiang Z (2007) Aromatic poly(ether ketone)s with pendant sulfonic acid phenyl groups prepared by a mild sulfonation method for proton exchange membranes. Macromolecules 40:1934–1944
67. Liu B, Robertson GP, Kim D-S, Sun X, Jiang Z, Guiver MD (2010) Enhanced thermo-oxidative stability of sulfophenylated poly(ether sulfone)s. Polymer 51:403–413
68. Pang J, Zhang H, Li X, Jiang Z (2007) Novel wholly aromatic sulfonated poly(arylene ether) copolymers containing sulfonic acid groups on the pendants for proton exchange membrane materials. Macromolecules 40:9435–9442
69. Pang J, Zhang H, Li X, Liu B, Jiang Z (2008) Poly(arylene ether)s with pendant sulfoalkoxy groups prepared by direct copolymerization method for proton exchange membranes. J Power Sources 184:1–8
70. Pang J, Zhang H, Li X, Ren D, Jiang Z (2007) Low water swelling and high proton conducting sulfonated poly(arylene ether) with pendant sulfoalkyl groups for proton exchange membranes. Macromol Rapid Commun 28:2332–2338
71. Pang J, Zhang H, Li X, Wang L, Liu B, Jiang Z (2008) Synthesis and characterization of sulfonated poly(arylene ether)s with sulfoalkyl pendant groups for proton exchange membranes. J Membr Sci 318:271–279
72. Lee JK, Li W, Manthiram A (2009) Poly(arylene ether sulfone)s containing pendant sulfonic acid groups as membrane materials for direct methanol fuel cells. J Membr Sci 330:73–79

73. Lafitte B, Jannasch P (2007) Proton-conducting aromatic polymers carrying hypersulfonated side chains for fuel cell applications. Adv Funct Mater 17:2823–2834
74. Parvole J, Jannasch P (2008) Poly(arylene ether sulfone)s with phosphonic acid and bis(phosphonic acid) on short alkyl side chains for proton-exchange membranes. J Mater Chem 18:5547–5556
75. Parvole J, Jannasch P (2008) Polysulfones grafted with poly (vinylphosphonic acid) for highly proton conducting fuel cell membranes in the hydrated and nominally dry state. Macromolecules 41:3893–3903
76. Kim DS, Kim YS, Guiver MD, Pivovar BS (2008) High performance nitrile copolymers for polymer electrolyte membrane fuel cells. J Membr Sci 321:199–208
77. Kim DS, Kim YS, Guiver MD, Ding J, Pivovar BS (2008) Highly fluorinated comb-shaped copolymer as proton exchange membranes (PEMs): fuel cell performance. J Power Sources 182:100–105
78. Kreuer KD (2001) On the development of proton conducting polymer membranes for hydrogen and methanol fuel cells. J Membr Sci 185:29–39
79. Yoshimura K, Iwasaki K (2009) Aromatic polymer with pendant perfluoroalkyl sulfonic acid for fuel cell applications. Macromolecules 42:9302–9306
80. Mikami T, Miyatake K, Watanabe M (2010) Poly(arylene ether)s containing superacid groups as proton exchange membranes. ACS Appl Mater Interfaces 2:1714–1721
81. Miyatake K, Shimura T, Mikami T, Watanabe M (2009) Aromatic ionomers with superacid groups. Chem Commun 42:6403–6405
82. Hirakimoto T, Fukushima K, Li Y, Takizawa S, Hinokuma K, Senoo T (2008) Fullerene-based proton-conductive material for the electrolyte membrane and electrode of a direct methanol fuel cell. ECS Trans 16:2067–2072
83. Watanabe M, Uchida H, Emori M (1998) Analyses of self-humidification and suppression of gas crossover in Pt-dispersed polymer electrolyte membranes for fuel cells. J Electrochem Soc 145:1137–1141
84. Watanabe M, Uchida H, Emori M (1998) Polymer electrolyte membranes incorporated with nanometer-size particles of pt and/or metal-oxides: experimental analysis of the self-humidification and suppression of gas-crossover in fuel cells. J Phys Chem B 102:3129–3137
85. Watanabe M, Uchida H, Seki Y, Emori M, Stonehart P (1996) Self-humidifying polymer electrolyte membranes for fuel cells. J Electrochem Soc 143:3847–3852
86. Dyck A, Fritsch D, Nunes SP (2002) Proton-conductive membranes of sulfonated polyphenylsulfone. J Appl Polym Sci 86:2820–2827
87. Gomes D, Buder I, Nunes SP (2006) Sulfonated silica-based electrolyte nanocomposite membranes. J Polym Sci Pol Chem 44:2278–2298
88. Karthikeyan CS, Nunes SP, Prado LASA, Ponce ML, Silva H, Ruffmann B, Schulte K (2005) Polymer nanocomposite membranes for DMFC application. J Membr Sci 254:139–146
89. Karthikeyan CS, Nunes SP, Schulte K (2005) Ionomer-silicates composite membranes: permeability and conductivity studies. Eur Polym J 41:1350–1356
90. Karthikeyan CS, Nunes SP, Schulte K (2006) Permeability and conductivity studies on ionomer-polysilsesquioxane hybrid materials. Macromol Chem Phys 207:336–341
91. Nunes SP, Ruffmann B, Rikowski E, Vetter S, Richau K (2002) Inorganic modification of proton conductive polymer membranes for direct methanol fuel cells. J Membr Sci 203:215–225
92. Silva VS, Ruffmann B, Silva H, Gallego YA, Mendes A, Madeira LM, Nunes SP (2005) Proton electrolyte membrane properties and direct methanol fuel cell performance. J Power Sources 140:34–40
93. Silva VS, Schirmer J, Reissner R, Ruffmann B, Silva H, Mendes A, Madeira LM, Nunes SP (2005) Proton electrolyte membrane properties and direct methanol fuel cell performance. J Power Sources 140:41–49
94. Su Y-H, Liu Y-L, Sun Y-M, Lai J-Y, Guiver MD, Gao Y (2006) Using silica nanoparticles for modifying sulfonated poly (phthalazinone ether ketone) membrane for direct methanol fuel cell: a significant improvement on cell performance. J Power Sources 155:111–117
95. Su Y-H, Liu Y-L, Sun Y-M, Lai J-Y, Wang D-M, Gao Y, Liu B, Guiver MD (2007) Proton exchange membranes modified with sulfonated silica nanoparticles for direct methanol fuel cells. J Membr Sci 296:21–28
96. Anilkumar GM, Nakazawa S, Okubo T, Yamaguchi T (2006) Proton conducting phosphated zirconia-sulfonated polyether sulfone nanohybrid electrolyte for low humidity, wide-temperature PEMFC operation. Electrochem Commun 8:133–136
97. Miyatake K, Tombe T, Chikashige Y, Uchida H, Watanabe M (2007) Enhanced proton conduction in polymer electrolyte membranes with acid- functionalized polysilsesquioxane. Angew Chem Int Ed 46:6646–6649

Books and Reviews

Colomban P (1992) Proton conductors: solids, membrane and gels – materials and devices. Cambridge University Press, Cambridge

Hickner MA, Ghassemi H, Kim YS, Einsla BR, McGrath JE (2004) Alternative polymer systems for proton exchange membranes (PEMs). Chem Rev 104:4587–4612

Kreuer KD, Paddison SJ, Spohr E, Schuster M (2004) Transport in proton conductors for fuel-cell applications: simulations, elementary reactions, and phenomenology. Chem Rev 104:4637–4678

Rikukawa M, Sanui K (2000) Proton-conducting polymer electrolyte membranes based on hydrocarbon polymers. Prog Polym Sci 25:1463–1502

Scherer GG (2008) Advances in polymer science: fuel cells I & II. Springer, Berlin

Tant BR, Mauritz KA, Wilkes GL (1997) Ionomers – synthesis, structure, properties and applications. Blackie Academic & Professional, New York

Mesoscopic Solar Cells

Michael Grätzel
Laboratory of Photonics and Interfaces, Institute of Chemical Science and Engineering, Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland

Article Outline

Glossary

AM1.5 Air mass 1.5: Defines position of the sun where the path through the atmosphere is 1.5 longer than at a vertical incidence.

Fill factor of the cell Maximum power output of the cell divided by the product of open circuit photovoltage (Voc) times the short circuit photocurrent density.

IPCE Incident photon to electric current conversion efficiency, presents the ratio of the electric current generated by monochromatic light of a certain wavelength over the incident photon flux.

Mesoscopic Size domain between 2 and 50 nm.

Power conversion efficiency (PCE) the maximal electric power generated by the photovoltaic cell divided by the incident solar light intensity under AM 1.5 standard reporting conditions (Intensity of the sunlight 1000W/m2 and T= 298 K).

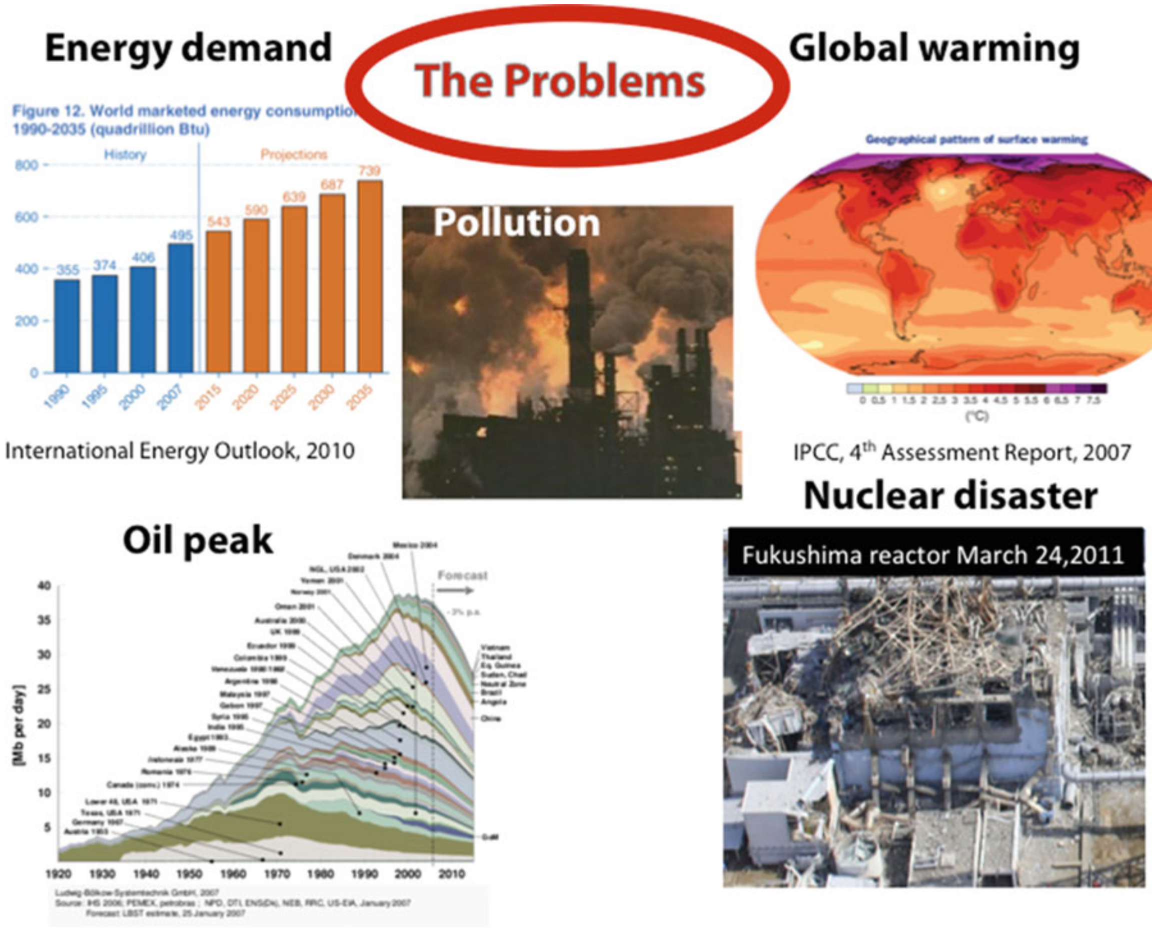
Sensitizer Dye molecule generating electric charges from sunlight.

Introduction

Perhaps, the largest challenge for our global society is to find ways to replace the slowly but inevitably vanishing fossil fuel supplies by renewable resources. The problem is compounded by an increase in the worldwide consumption of energy, which is expected to double within the next 40 years from the current level of 500 exajoules/year (exa = 10^{18}) to 1,000 exajoules/year. This additional demand cannot be met by accelerated combustion of fossil fuels, which would entail enhanced environmental pollution and global warming, leave alone the fact that oil production has already peaked and will decline in the future (Fig. 1). Furthermore, the current ongoing disaster at the Fukushima reactor site in Japan along with previous major accidents has exposed to the world the risks and limitations of nuclear energy use, leave alone that the issue of where to store nuclear waste over ten thousands of years in a safe manner and at what cost remains unresolved to date.

The sun provides about 120,000 TW to the earth's surface which amounts to 6,000 times the present rate of the world's energy consumption. However, capturing solar energy and converting it to heat, electricity, or chemical fuels, such as hydrogen, at low cost and using abundantly available raw materials remain a huge challenge. Photovoltaic cells are expected to make pivotal contributions to identify environmentally friendly solutions to this energy problem. One area of great promise is that of a new generation of nonconventional photovoltaic converters generally referred to as mesoscopic solar cells (MSCs), which employ three-dimensional (bulk) interpenetrating network junctions for light harvesting and afford charge carrier transport using structural elements with features in the 2–50-nm size range.

While this field is still in its infancy, it is presently receiving enormous attention, thousands of publications having appeared over the last decade and dozens of industrial enterprises being involved in the commercialization of this new technology. The advantage of the mesoscopic solar cells is that they can be produced at low cost, i.e., potentially significantly less than 1 US$/peak watt. These systems have the potential to deliver solar electricity at a price of 5 cents/KWh, which is competitive with present conventional energy cost. Some, but not all MSC embodiments, can avoid the expensive and energy-intensive high vacuum and



Mesoscopic Solar Cells. Figure 1
Projected growth of future energy demand and problems arising from the use of fossil fuels or nuclear reactors to cover the additional needs of 500 exajoules by 2050

materials purification steps that are currently employed in the fabrication of all other thin film solar cells. They employ materials that are abundantly available so that the technology can be scaled up to the terawatt level without running into feedstock supply problems. This gives mesoscopic solar cells an advantage over the two major competing thin film photovoltaic devices, i.e., CdTe and Cu(In, Ga)Se_2 (CIGS), which are based on highly toxic materials or elements of low natural abundance. While some MSCs do use rare elements such as Ru and Pt the quantities employed are so small that these noble metals are not cost determining and do not limit the scale up of the technology to the terawatt level. However, a drawback of the current embodiment of MSCs is that their efficiency is still lower than that for single and multicrystalline silicon as well as CdTe and Cu(In, Ga)Se_2 (CIGS) cells. Also, MSCs based on conjugated organic compounds, such as blends of C60 with polythiophenes, are very sensitive to water and oxygen and, hence, need to be carefully sealed to avoid rapid degradation. The present entry focuses on dye-sensitized solar cells (DSCs), which are leading this new generation of mesoscopic photovoltaic devices [1]. Research in the DSC field is booming, and the technology is advancing at a very rapid pace. Progress in several important areas related to the sensitization of mesoscopic semiconducting oxides by dyes and quantum dots has been covered by a number of recent reviews and a book [2–23]. Here, we discuss the operational principles of the device and some recent exciting developments stemming mainly from our own laboratory. We present promising new concepts to boost further the conversion efficiency and stability of the DSC and summarize the status of its commercial applications.

Operational Principles of Dye-Sensitized Mesoscopic Solar Cells

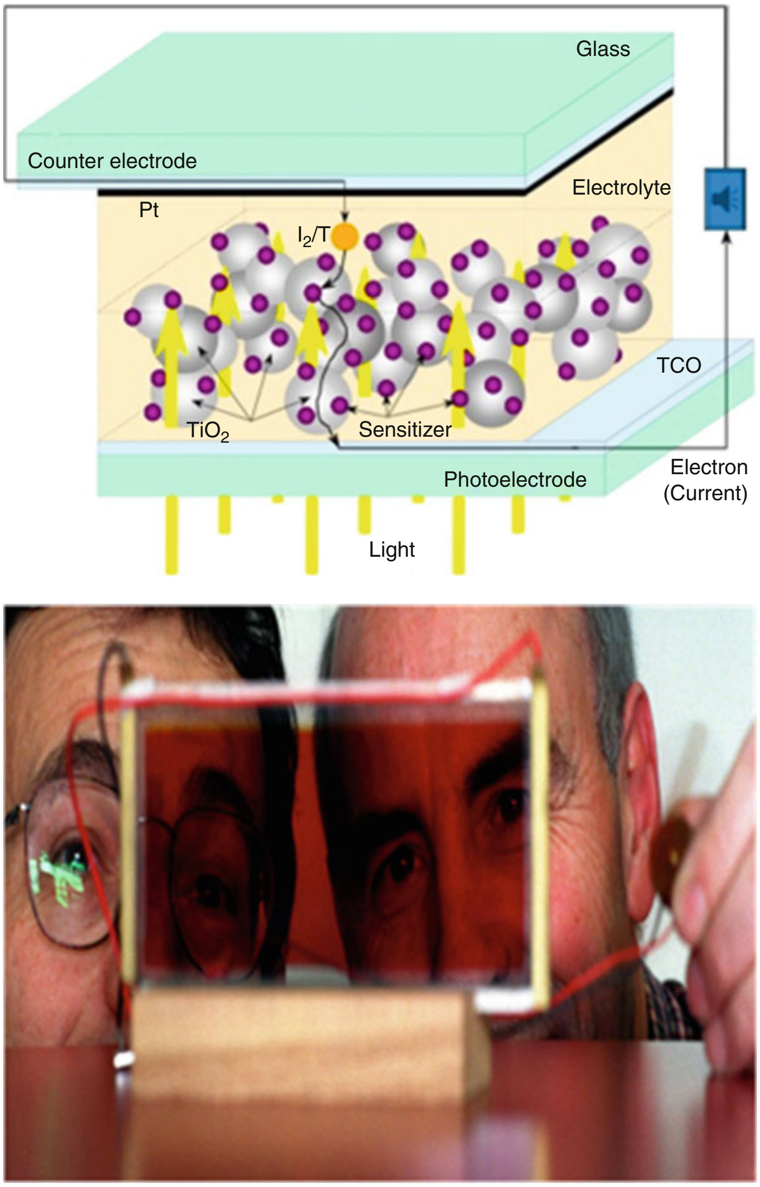
Shown in the upper part of Fig. 2 is the cross section of a typical embodiment for a mesoscopic solar cell based on the sensitization of nanocrystalline titania (TiO_2, anatase) films. The nanoparticles (shown as grey balls) are screen-printed onto a glass sheet covered by a thin film of a transparent conducting oxide (TCO). The latter collects the electrons generated by light excitation of the sensitizer molecules (shown as red dots). The excited state of the sensitizer injects electrons in the conduction band of the TiO_2 particles. The conduction band electrons percolate subsequently across the nanoparticle network, before they reach the TCO front contact. From there, they pass through the external circuit to perform electrical work and reenter the cell through the counter electrode. A redox shuttle, typically the iodide/iodide couple, moves the electrons back to the surface of the nanocrystalline film where it regenerates the sensitizer.

The lower part of Fig. 2 shows a photograph of a dye-sensitized solar cell. The transparent glass module produces electricity from ambient light captured from all spatial angles due to its bifacial character, which is used to drive a fan. The photograph illustrates the transparent character of electricity generating glass which is a unique feature of dye-sensitized mesoscopic solar cells, offering widespread applications as electric power producing window in glass facades of buildings.

Following its inception in 1985 [24], the dye-sensitized mesoscopic solar cells (DSC) were the first device to use a three-dimensional interpenetrating network junction, referred to as "bulk heterojunction" for solar light energy harvesting and conversion. Leading a new generation of solar cells [24–31], the DSC is the only photovoltaic device that uses molecules or semiconductor quantum dots to absorb photons and convert them to electric charges without the need of intermolecular transport of electronic excitation. The conversion of light to electricity does not involve the participation of minority carriers. The role of the sensitizer or quantum dot is to absorb light and generate positive and negative electric charges, which are injected in appropriate charge transport materials, i.e., an n-type conductor such as TiO_2 for the electrons and a p-type conductor or electrolyte for the positive charges (holes).

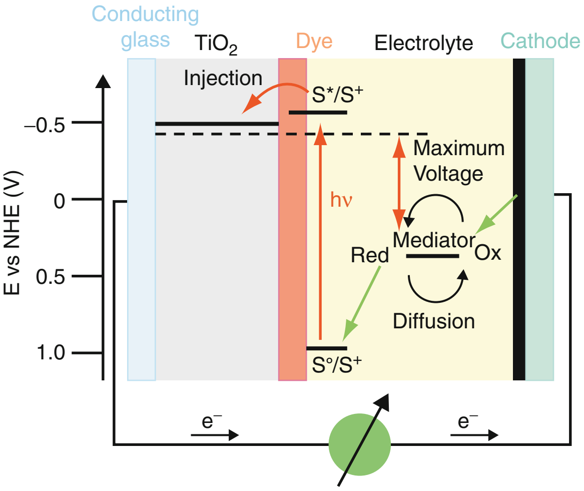
Figure 3 shows a typical energy band diagram of the DSC. Sunlight is harvested by the sensitizer that is attached to the surface of a large-bandgap semiconductor, typically a film of TiO_2, ZnO, or SnO_2 constituted of nanoparticles, nanorods, nanotubes, or other mesoscopic structures. Photoexcitation of the dye results in the injection of electrons into the conduction band of the oxide, producing the oxidized form of the sensitizer S^+. The dye is regenerated by electron donation from an organic or inorganic hole conductor or a redox electrolyte that is infiltrated into the porous films. Figure 3 shows two typical embodiments of the DSC. The left drawing refers to the version that employs a redox electrolyte while the right configuration employs a solid state hole conductor to shuttle the positive charges generated under light excitation from the oxidized sensitizer to the back contact. The former contains, most frequently, the iodide/triiodide couple as a redox shuttle, although other mediators such as cobalt (II/III) complexes [30–32], the TEMPO/TEMPO + redox couple [33] thiotetrazol [34] tetramethylthiourea [35], and ferrocene [36] have shown promise recently as an alternative to the $I^-/I_3{}^-$ system. Electron transfer from the reduced mediator to S^+ regenerates the original form of the dye while producing the oxidized form of the mediator. This prevents any significant buildup of S^+ at the surface, which could recapture the conduction band electron. The reduced mediator is regenerated in turn at the counter electrode. Electrons are supplied via migration through the external load completing the cycle. Thus, the device is regenerative producing electricity from light without any permanent chemical transformation. Figure 4 shows two typical embodiments of the DSC using a redox electrolyte or solid hole conductor for charge transport between the working and counter-electrode.

The open-circuit photovoltage V_{oc} produced under illumination corresponds to the difference between the chemical potential (Fermi level), $\mu(e^-)$, attained by the conduction band electrons in the mesoscopic titania



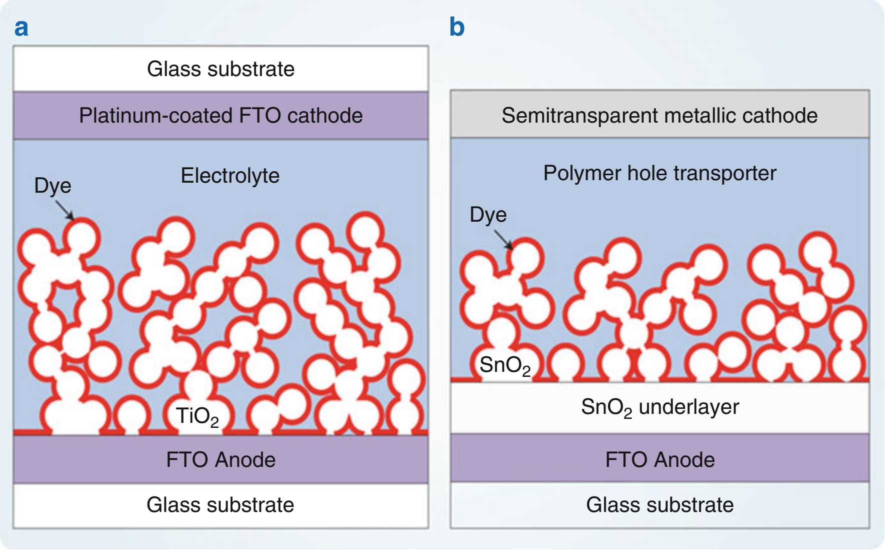
Mesoscopic Solar Cells. Figure 2

Upper part: Cross section of a typical embodiment for a mesoscopic solar cell based on the sensitization of nanocrystalline titania (TiO_2, anatase) films. The nanoparticles (shown as *grey balls*) are screen-printed onto a glass sheet covered by a thin film of a transparent conducting oxide (TCO). The latter collects the electrons generated by light excitation of the sensitizer molecules (shown as *red dots*). The excited state of the sensitizer injects electrons in the conduction band of the TiO_2 particles. The conduction band electrons percolate subsequently across the nanoparticle network, before the reach the TCO front contact. From there, they pass through the external circuit to perform electrical work and reenter the cell through the counter electrode. A redox shuttle, typically the iodide/iodide couple, moves the electrons back to the surface of the nanocrystalline film where the regenerate the sensitizer. *Lower part*: Photograph of a dye-sensitized solar cell. The transparent glass module produces electricity from ambient light from all angles due to its bifacial character, which is used to drive a fan. The photograph illustrates the transparent character of glass which is a unique feature of dye-sensitized mesoscopic solar cells, offering widespread applications as electric power producing window in glass facades of buildings film under illumination and the chemical potential $\mu(h^+)$ of the holes in the hole conductor. For liquid or solid electrolytes, the latter corresponds to the Nernst potential (E) of the redox couple used as a shuttle to transport positive charges from the sensitizer to the back contact of the cell. Note that in the dark at equilibrium $\mu(e^-) = \mu(h^+)$, i.e., the Fermi level is constant within the whole device.



Mesoscopic Solar Cells. Figure 3
Energy band diagram of a typical embodiment of the DSC. The energy levels are presented on the normal hydrogen reference electrode scale for N719 as sensitizer and iodide/triiodide as redox electrolyte

The energy levels in Fig. 3 are drawn to fit a frequently employed embodiment of the DSC based on the N719 ruthenium dye (the di-tetrabutylammonium salt of N3), the iodide/triiodide as a redox couple, and nanocrystalline anatase films as electron collector. The ground state standard redox potential of the N719 as well as that of several other ruthenium complexes and organic sensitizers is around 1 V against the normal hydrogen electrode (NHE) when it is adsorbed at the titania surface, while the Nernst potential of the triiodide/iodide-based redox electrolyte is between 0.3 V and 0.4 V, and the energy of the conduction band edge of anatase is located at −0.5 V. Neglecting entropy changes during the light absorption and using a 0–0 vibronic transition energy of 1.65 eV, the excited state redox potential of N719 is derived to be −0.65 eV. Hence, the driving force for electron injection under standard conditions is 0.15 eV. By contrast, the regeneration consumes an excessive amount of free energy, i.e., between 0.6 and 0.7 eV depending on the exact composition of the iodide-/triiodide-based electrolyte.



Mesoscopic Solar Cells. Figure 4
Two typical embodiments of dye-sensitized mesoscopic solar cells. (**a**) The dye-loaded mesoscopic TiO$_2$ film is contacted by a redox electrolyte whose task is to transport positive charges from the sensitized to the counter electrode. (**b**) A polymer or organic molecular hole conductor assumes the role of transporting the positive charge

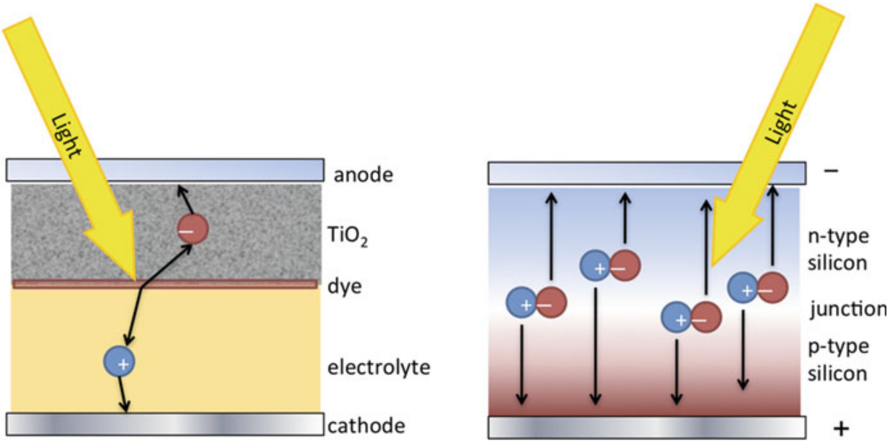
This mismatch of over 0.6 eV between the redox level of the sensitizer and that of the triiodide/iodide presents the major loss channel for DSCs using this type of electrolyte. Using the black dye (N749) instead of N719, the loss is reduced to 0.4 eV due to its 0.2 eV lower ground-state redox potential [37]. Measurements of a series of ruthenium complexes have shown that the minimum driving force required for the near quantitative interception of the geminate electron recombination with the oxidized sensitizer by oxidation of iodide to triiodide is about 0.3 eV. This relatively high value results from the two-electron character of the iodide oxidation reaction, which passes through I_2^- radicals as intermediates. A recent breakthrough in the DSC field opened up the way to use one electron redox mediators, reducing the losses during the regeneration of the sensitizer to merely 0.2 eV. This has allowed to increase the Voc value of DSCs to over 1 V.

Comparison with Conventional p–n Junction Devices

One of the fundamental advantages of the DSC is that it separates the two functions of light harvesting and charge carrier transport, whereas conventional silicon-based cells and all known organic photovoltaic devices perform both operations simultaneously. This imposes stringent demands on the optical and electronic properties of the semiconductor, i.e., its bandgap and conduction band or valence band position, as well as mobility and the recombination time of photogenerated charge carriers. As a result, the choice of suitable materials that are able to act as efficient photovoltaic converters is greatly restricted. The separation of the functions of light harvesting and carrier transport on the other hand opens up many options for the absorber and charge transport materials. The molecular sensitizer or semiconductor quantum dot is placed at the interface between an electron (n) and hole (p) conducting material (Fig. 5 left side). The former is typically a wide band semiconductor oxide, such as TiO_2, ZnO, or SnO_2, while the latter is a redox electrolyte or a p-type semiconductor. Upon photoexcitation, the sensitizer injects an electron in the conduction band of the oxide and is regenerated by hole injection in the electrolyte or p-type conductor. Alternatively, the excited sensitizer may inject a hole in the valence band of p-type oxide such as NiO [32] and is regenerated from its reduced form by electron transfer to an acceptor in the electrolyte. In both cases, the role of the sensitizer is to absorb light and generate positive and negative charge carriers. The latter are transported to the front and back contacts where they are collected as electric current. The open-circuit photovoltage V_{oc} corresponds to the difference in the Fermi level of the n- and p-type conductor under illumination. This configuration has the following advantages:

- Constraints on individual components of the cell are relaxed, i.e., the electron and hole conductor materials, as well as the sensitizer or quantum dot type, can each be separately selected and tuned for optimal performance.
- Light absorption and charge transport are decoupled avoiding the participation of minority carriers in the light energy conversion process.
- The optoelectronic properties of the mesoscopic semiconductor oxide film, which supports the self-assembled dye monolayer, can be adapted to yield efficient light harvesting and charge carrier collection. Apart from exploiting light scattering and photon containment effects, structures that display photonic bandgaps and plasmonic behavior have shown benefits. These light management strategies allow reducing the oxide film thickness and hence the amount of substances required to harvest solar light, saving materials cost.
- The use of a molecular sensitizer avoids the need for transport of excitons within the absorber material to achieve photoinduced charge separation.
- Finally, the sensitizer itself if present as a well-organized compact monomolecular layer may retard the recombination of photogenerated charge carriers, which has to occur across the interface occupied by the dye molecules.

For comparison, the right side of Fig. 5 shows a schematic illustration of the operational principle for a conventional silicon p–n junction photovoltaic cell. Here, the same semiconductor material, e.g., silicon, assumes the two key functions of light absorption and charge carrier transport. Photoexcitation of the silicon generates electron–hole pairs which need to reach the p–n junction before they recombine.



Mesoscopic Solar Cells. Figure 5
Left side: Schematic illustration of the operation principle for dye-sensitized photovoltaic cell. The heart of the device is a sensitizer or quantum dot placed at the interface between an electron (n)-conducting wide bandgap semiconductor material, such as TiO_2, and a hole (p) conductor or electrolyte. Upon photoexcitation, it injects an electron in the conduction band of a wide bandgap semiconducting oxide and is regenerated by hole injection in the redox electrolyte or p-type conductor. The charges diffuse to the front and back contacts where they are collected as electric current. The open-circuit photovoltage corresponds to the difference in the Fermi level of the n- and p-type conductor under illumination. *Right side*: Schematic illustration of the operation principle for conventional silicon p–n junction photovoltaic cell. The same semiconductor material assumes the two key functions of light absorption and charge carrier transport. Photoexcitation of the silicon generates electron–hole pairs which need to reach the p–n junction before they recombine. The local electric field present in the junction separates the charges attracting electrons to the n-doped and holes to the p-doped layer. The diffusion length of the minority carriers, i.e., of the electrons and holes in the p- and n-doped silicon layer, respectively, must be at least ten times larger than the layer thickness to collect more than 99% of the photogenerated charge carriers. This requires precise doping as well as high crystallinity and low levels of defects and impurities imposing stringent conditions on the purity of the semiconductor (>99,999% for silicon)

The local electric field present in the junction separates the charges, attracting electrons to the n-doped and holes to the p-doped layer. The diffusion length of the minority carriers, i.e., that of the electrons and holes in the p- and n-doped silicon layer, respectively, must be at least ten times larger than the layer thickness to collect more than 99% of the photogenerated charge carriers. This requires precise doping as well as a high crystallinity and a low level of defects and impurities imposing stringent conditions on the purity of the semiconductor (>99.9999% for solar grade silicon).

Note that minority carriers, i.e., electrons and positive charges (holes) in p- and n-doped semiconductors, respectively, play no role in the photovoltaic conversion process accomplished by the DSC. Because the sensitizer injects electrons into the n-type and holes into the p-type collector, only majority carriers are generated. These charges move in their respective transport medium to the front and back contacts of the photocell where they are collected as electric current. Therefore, their recombination has to occur across the interface between the electron and hole conducting material.

By contrast, in conventional p–n junction photovoltaic cells, the photocurrent arises from minority carriers generated by the photoexcitation of the semiconductor. The minority carrier must live long enough to reach the junction formed between the p- and n-doped semiconductor material before recombination with the majority carriers takes place. The electric field present in the vicinity of the junction separates the positive and negative charges generated under illumination attracting the electrons to the n-doped and the holes to the p-doped material. In order to impart a sufficiently long lifetime to the photogenerated electron–hole pairs, the use of very pure materials is required. The chemical purification of the semiconductor contributes largely to the high cost and long energy payback time of silicon photovoltaic cells.

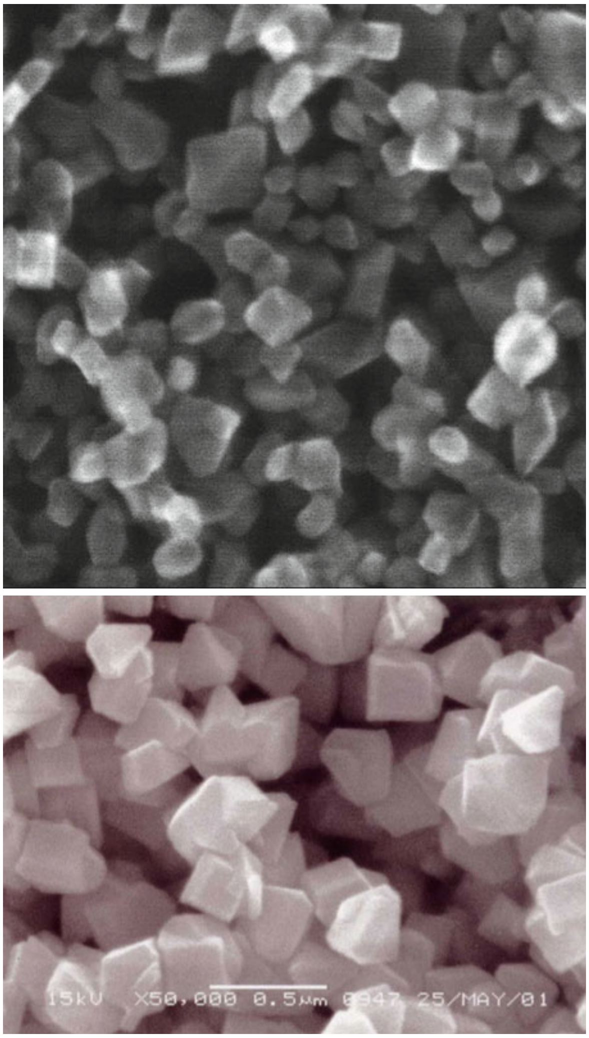
In the DSC, the recombination of charge carriers occurs across the phase boundary separating the electron from the hole conductor medium. This inherent geometry offers the unique prospective to fashion the interface in a judicious manner in order to retard the back electron transfer reaction. One promising approach to accomplish this goal is the molecular engineering of sensitizers forming a self-assembled compact monolayer alone or in conjunction with a coadsorbent at the oxide surface. Such an insulating film would impair the backflow of electrons across the junction reducing the back reaction rate and increasing the overall solar to electric power conversion efficiency of the cell.

The Virtues of the Nanostructure

Figure 6 shows on the top a scanning electron microscopy picture of a mesoscopic TiO_2 (anatase) layer. The particles have an average size of 20 nm, and the facets exposed have mainly (101) orientation, corresponding to the anatase crystal planes with the lowest surface energy (ca. 0.5 J/m^2).

The light harvesting by the surface-adsorbed sensitizer can be further improved by introducing larger titania particles in the film that scatter light. These are either mixed with or printed on top of the film of 15–30-nm-sized TiO_2 nanoparticles. The scattered photons are contained in the film by multiple reflections increasing their optical path length substantially beyond the film thickness. As a result, the absorption of solar light is enhanced, particularly in the red and near IR spectral region where the currently used ruthenium complexes show only weak light absorption. For example, using 200–400-nm-sized anatase particles as light scattering centers increases the Jsc of N719-based DCCs by as much as 3–4 mA/cm^2 due to the enhanced absorption of red or near infrared photons [38]. A scanning electron micrograph of such particles is shown on the bottom of Fig. 6.

A whole range of nanostructures have been tested so far ranging from simple assemblies of nanoparticles to organized mesoporous films [39] nanorods [40] and

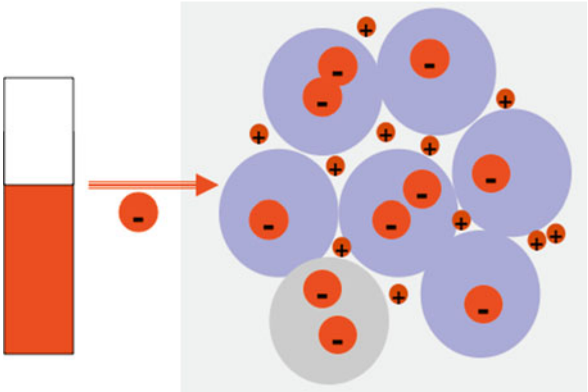


Mesoscopic Solar Cells. Figure 6
Top: Scanning electron microscope picture of a transparent nanocrystalline TiO_2 film formed by ca. 20-nm-sized anatase particles. Note the *bipyramidal* shape of the particles having (101) oriented facets exposed. *Bottom*: Scanning electron microscope image of 200–400-nm-sized anatase particles, which are employed as light scattering centers. The larger particles are either printed as a film on the transparent particles or mixed directly with the 20-nm-sized TiO_2 colloid to augment the optical path by a multiple light scattering effect improving the absorption of sunlight in particular in the red and near IR spectral region where the light absorption by the sensitizer is weak. Scale bars are 200 nm and 500 nm for the *top* and *bottom* SEM photographs, respectively

nanotubes [41, 42]. These studies are motivated by the expectation that the transport of charge carriers along the tubes is more facile than within a random network of nanoparticles where the electrons have to cross many particle boundaries. Hence, one-dimensional nanostructures should produce a lower diffusion resistance than the nanocrystalline films facilitating the collection of photogenerated charge carriers.

The mesoscopic morphology of the oxide semiconductor film is crucial for the efficient operation of a DSC. The key advantages are as follows:

- The mesoscopic structure allows to harvest sunlight efficiently even by a single layer of surface-adsorbed sensitizer. On a flat surface, a monolayer of dye absorbs at most a few percent of the impinging light because it occupies an area that is several hundred times larger than its optical cross section. Employing a mesoscopic structure to support the sensitizer overcomes this notorious inefficiency problem. The real surface area of a 10-μm-thick film formed by 20-nm-sized anatase particles is at least 1,200 times larger than the projected area. This increases dramatically the light-harvesting capacity by the adsorbed monolayer, which for state-of-the-art panchromatic sensitizers is practically quantitative over the whole visible and near IR range.
- The use of oxide nanoparticles to transport the electrons injected by the photoexcited sensitizer into their conduction band avoids space-charge control of the photocurrents as the charge injected is screened by positive ions present in the electrolyte or at the surface of the oxide (Fig. 7). This greatly facilitates electron percolation across the film. The electron charge is screened by the cations present in the electrolyte or on the particle surface, which eliminates the internal field, so no drift term appears in the transport equation.
- The mesoscopic large-bandgap semiconductor oxide films are insulating in the dark. This is an advantage as the presence of significant levels of electrons in the conduction band of the oxide would lead to Auger-type energy transfer quenching of the sensitizer-excited state, reducing the quantum yield of carrier generation. However, under light, the conductivity of the films augments by several orders of magnitude. In fact, a single electron injected in a 20-nm-sized particle produces an electron concentration of 2.4 × 10^{17} cm^{-3}. This corresponds to a specific conductivity of 1.6 × 10^{-2} S cm^{-1} if a value of 10^{-4} cm^2 s^{-1} is used for the electron diffusion coefficient. In reality, the situation is more complex because the transport of charge carriers in these films involves trapping unless the Fermi level of the electron is so close to the conduction band that all of the traps are filled and the electrons are moving freely. Therefore, the depth of the traps that participate in the electron motion affects the value of the diffusion coefficient. This explains the observation that the diffusion coefficient increases with the light intensity. Random walk modeling gives an excellent description of the intricacies of the electron transport in such mesoscopic semiconductor films [43].



Mesoscopic Solar Cells. Figure 7
Electron transport in nanocrystalline TiO_2 films. Space-charge control of the photocurrent is avoided by screening of the negative charge by the cations present in the electrolyte or at the particle surface

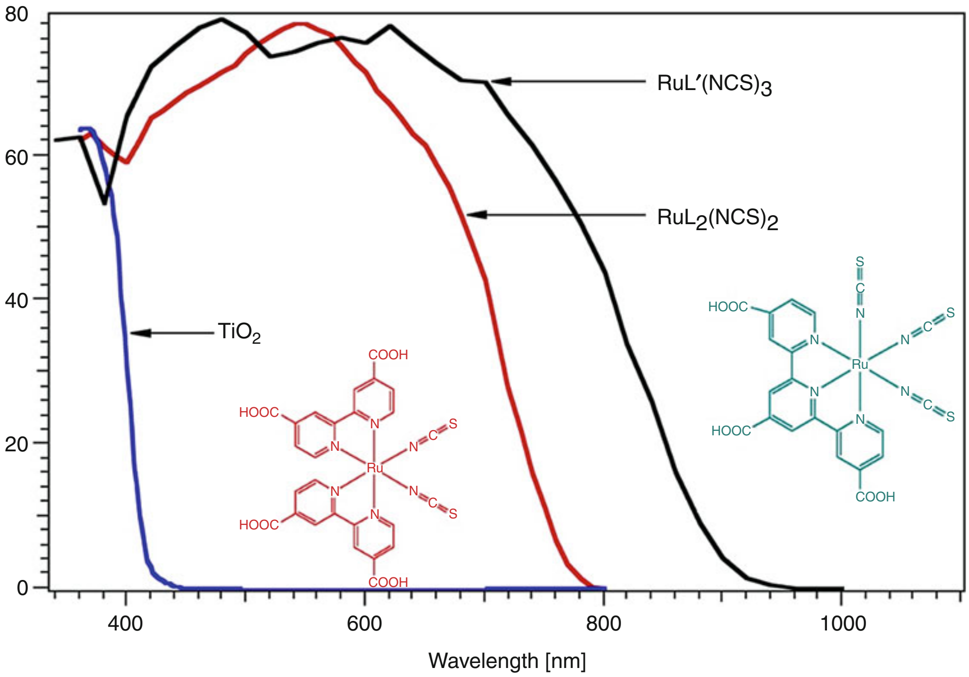
Solids containing periodic pore structures that exhibit photonic bandgaps show also great promise for enhancing the red response of the DSC [44]. The benefits from using such photon capture strategies are unquestionable, as they have been shown to enhance the photocurrent response of the DSC in particular in the near IR and red region of the solar spectrum. The gain in short-circuit photocurrent and overall conversion efficiency achieved by exploiting these optical effects can be as high as 30%.

Typical Sensitizers Based on Ruthenium Complexes

The chemical structures of two typical sensitizers, i.e., the N3 and the black dye (N749), are shown in Fig. 8 together with spectral response of the photocurrent generated by these sensitizers. The incident photon to current conversion efficiency (IPCE), sometimes referred to also as the "external quantum efficiency" (EQE), is plotted as a function of excitation wavelength. The IPCE value corresponds to the photocurrent density produced in the external circuit under monochromatic illumination of the cell divided by the photon flux that strikes the cell. The following product expresses this key parameter:

$$\mathrm{IPCE}(\lambda) = \mathrm{LHE}(\lambda)\phi_{\mathrm{inj}}\ \eta_{\mathrm{coll}} \qquad (1)$$

Here, LHE(λ) is the light-harvesting efficiency for photons of wavelength λ, ϕ_{inj} is the quantum yield for electron injection from the excited sensitizer in the conduction band of the semiconductor oxide, and η_{coll} is the electron collection efficiency. Figure 8 shows that the IPCE reaches close to 80% for both sensitizers in the plateau region of the IPCE spectrum. Compared to the N3 sensitizer, the spectral response of the black dye extends further into the near IR, its photocurrent onset being around 900 nm instead of 800 nm for the former dye. This corresponds to a bandgap of 1.4 eV, which is close to the optimum threshold absorption for single-junction photovoltaic cells. However, for both sensitizers, the IPCE curve rises only gradually from the absorption onset to shorter wavelengths due to the low extinction coefficients of the sensitizers in the longer wavelength range. This behavior results in a very significant loss in photocurrent. For example, the photocurrent density obtained currently with the N749 (black) dye



L = 4,4′-COOH-2,2′-bipyridine
L′ = 4,4′,4″-COOH-2,2′:6′,2″-terpyridine

Mesoscopic Solar Cells. Figure 8
Incident photon to current conversion efficiency as a function of wavelength for the standard ruthenium sensitizers N3 (*red line*), the black dye N749 (*black curve*), and the blank nanocrystalline TiO_2 film (*blue curve*). The chemical structure of the sensitizers is shown as *insets*

under standard conditions is close [37, 45] to 21 mA/cm^2, while the maximum J_{sc} for a sensitizer with an absorption onset of 900 nm is 33 mA/cm^2. Even if one concedes an IPCE loss of 10% across the whole spectral absorption domain of the black dye, reducing Jsc to 30 mA/cm^2, the gain in photocurrent would still boost the overall conversion efficiency of the device by a factor of 1.5 to over 15%. Improving the light harvesting in the 650–900-nm domain is therefore one of the greatest challenges faced by present day research in the DSC field. Even moderate improvements in the photoresponse of the sensitizer in red and near IR wavelength region will greatly benefit the conversion efficiency. An amazing advance in this area was recently announced by the Segawa group at the University of Tokyo. By replacing the 3 thiocyanate groups in the black dye with 2 chloride ions and one phosphine ligand a new panchromatic sensitizer, coded DX1, was fashioned that harvests sunlight over the whole visible and near IR range up to 1000 nm generating a short circuit photocurrent close to 27 mA/cm^2 [46].

The Present Status of Dye-Sensitized Solar Cells

The present state of the art of the dye-sensitized solar technology is summarized below in Table 1.

The overall conversion efficiency of the dye-sensitized cell is determined by the photocurrent density measured at short circuit (J_{sc}), the open-circuit photovoltage (V_{oc}), the fill factor of the cell (FF), and the intensity of the incident light (I_s).

$$\eta_{global} = J_{sc} \times V_{oc} \times FF/I_s \tag{2}$$

The fill factor can assume values between 0 and 1 and is defined by the ratio of the maximum power (P_{max}) of the solar cell divided by the open-circuit voltage (V_{oc}) and the short-circuit current (I_{sc}):

$$FF = P_{max}/(I_{sc} \times V_{oc}) \tag{3}$$

P_{max} is the product of photocurrent and photovoltage at the voltage where the cell's power output is maximal. The value of the fill factor reflects the extent of electrical (Ohmic) and electrochemical (overvoltage) losses occurring during operation of the DSC. Increasing the shunt resistance and decreasing the series resistance as well as reducing the overvoltage for diffusion and electron transfer will lead to higher fill factors, thus resulting in greater efficiency and pushing the cells output power closer toward its theoretical maximum.

Under full sunlight corresponding to the standard air mass 1.5 global (AM 1.5 G) spectral distribution, at an intensity I_s = 1,000 W/cm^2, J_{sc} values ranging from 16 to 22 mA/cm^2 are reached with state-of-the-art ruthenium sensitizers, while the V_{oc} attains 0.7–0.86 V, and typical values for the fill factor are 0.65–0.8. Note that the fill factor of a DSC is affected by the transfer coefficient β for the electron transfer from TiO_2 to the electrolyte, whose rate shows exponential voltage dependence, following a Tafel law [47]. For a given Voc, the larger the β value, the better the fill factor. Theoretically, β should be 1 for electron transfer from a semiconductor to a solution redox couple, but smaller values are usually observed due to the participation of surface states in the interfacial charge transfer, whose energy levels lie below the conduction band of $TiO_{2.}$ A larger β value can be associated with a shallower distribution of these defect states [47].

Until recently, the photovoltaic performance of the black dye measured under full AM 1.5 sunlight was superior to all other known charge-transfer sensitizers. A certified overall power conversion efficiency of 10.4% was reached with N749 already in year 2001 [37]. By 2006, a higher certified efficiency of 11.1% was attained using the same dye [45]. Recently the PCE has been

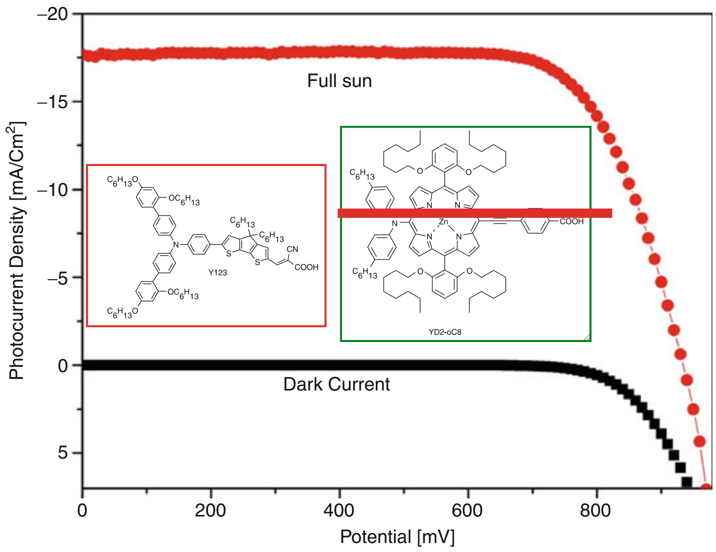
Mesoscopic Solar Cells. Table 1 Present state of the development of dye-sensitized solar cells. Note the small loss of efficiency encountered on going from a small laboratory cell to an industrially produced module

• *Power conversion efficiency* (PCE) measured under AM 1.5 sunlight (STC): laboratory cells: 12.3% [49], modules: 9.9% [54], tandem cells: 15–16% [50]
• *Stability* >20 years outdoors [59]
• *Energy payback time*: <1 year (3GSolar [60] and ECN [61] life cycle analysis)
• *Industrial development*: has been launched by several industrial companies, mass production of light weight flexible modules started in 2009 by G24Innovation (www.g24i.com)

further improved to 11.4% by the same group using a co-sensitizer along with the black dye [48]. Figure 8 shows that the spectral response of the photocurrent of the black dye is red-shifted by 100 nm with respect to that of the standard N3 dye, resulting in significantly higher short-circuit photocurrents, even though its surface coverage at monolayer saturation and its extinction coefficient are about 30% lower than the respective values obtained for the N3 dye [45].

However, recently new porphyrin sensitizers have been developed, in particular, the YD2-*o*-C8 zinc complex whose PCE exceeds that of the black dye when used in conjunction with Co(II/III)(bipy)$_3$ complexes as redox couple [49]. The Nernst potential of this redox mediator is more positive than that of the iodide/triiodide couple yielding substantial gains in open circuit potential of the cells. The performance of YD2-*o*-C8 is enhanced by judicious substitution of meso-positions at the tetrapyrrol ring with donor and acceptor moieties and the introduction of bulky alkoxy groups. This increases the light harvesting capacity of the porphyrin and blocks the unwanted interfacial electron back transfer. Further gain in efficiency was obtained via the use of a co-sensitizer coded Y123 absorbing strongly in the green spectral region. Figure 9 below shows a plot of the photocurrent density versus voltage for a mixture of the two dyes, co-adsorbed at the surface of the nanocrystalline titania film along with the structure of the sensitizers. The Jsc, Voc, and FF values measured under standard air mass (AM1.5) reporting conditions were 17.8 mA/cm^2, 0.935 V, and 0.74 yielding a power conversion efficiency of 12.3%.

The fact that DSCs are transparent and their optical properties are tunable by varying the type of sensitizer or the thickness of the nanocrystalline film renders them well suited for use in stacked tandem devices. While this development is still in its infancy, promising results have been obtained. Our previous work has shown that efficiencies in the 15–16 % range can be readily obtained in a stacked two-level tandem structure, employing a DSC on top of a CIGS film [50]. Spectral splitting of the solar light flux, e.g., by a heat mirror that passes visible light but reflects IR radiation is another attractive method to improve the efficiency of PV cells. Recent proof-of-concept results suggest that with DSCs as visible light and a silicon cell as IR-absorber system-level efficiencies approaching 20% should be achievable [51].



Mesoscopic Solar Cells. Figure 9

Photocurrent density vs voltage curve for a DSC employing the YD2-o-C8 porphyrin and Y123 as a sensitizer and co-sensitizer respectively. The structures of the two dyes are inserted in the diagram. The conversion efficiency achieved under standard reporting conditions, i.e. AM 1.5 sunlight of 1000W/m^2 intensity and 298K temperature is 12.3%

Reproducibility of Cell Fabrication and Scale-up

While the DSC can be produced in a relatively simple way in the laboratory without employing a glove box or high vacuum steps, a rigorous protocol needs to be applied during cell fabrication to achieve high efficiencies in a reproducible manner. By taking the appropriate precautions, relative variations of the efficiency of less than 2–3% can be readily achieved for laboratory cells. Thus, a detailed procedure providing a guide to realize reproducibly cell efficiency values over 10% has been published recently [52]. Reproducible manufacturing of DSC modules on a semiautomated baseline has also been reported [53]. Due to significant industrial upscaling efforts, the conversion efficiency of DSC modules has been steadily rising over the last few years, the certified value measurer under AM 1.5 standard conditions reaching currently 10% [54].

Stability and Commercial Development of the DSC

Long-term stability is a key requirement for all types of solar cell. A vast amount of tests have therefore been carried over the last 15 years to scrutinize the stability of the DSC both by academic and industrial institutions. Most of the earlier work has been reviewed [55, 56]. Long-term accelerated light-soaking experiments performed over many 1,000 h under full or even concentrated sunlight have confirmed the intrinsic stability of current DSC embodiments [57]. Stable operation under high temperature stress at 80–85°C as well as under damp heat and temperature cycling has been achieved by judicious molecular engineering of the sensitizer, the use of a robust and nonvolatile electrolytes such as ionic liquids [58] and adequate sealing materials. In the early development stage of the DSC technology, the quality of device sealing was sometimes not appropriate in laboratory test cells, causing leakage of the volatile nitrile-based solvents, typically used for the electrolytes. While this is occasionally still considered as a problem, most research groups with longer practical experience, including industrial enterprises, have overcome this by improving the sealing methods. Due to the direct relevance to the manufacturing of commercial products, little is published on these processing issues though. Good results on overall system endurance have been reported since several years demonstrating excellent stability under accelerated laboratory test conditions. DSC lifetimes of over 20 years have been projected from continuous light soaking tests performed over 20000 hours [59]. These promising results are presently being confirmed under real outdoor conditions. Thus the Israeli company 3GSolar recently announced that its DSC modules have been operating on the company's rooftop continuously for two years and they continue to perform to the same standard as on the day they were placed outdoors [60]. Importantly, 3GSolar [60] and ECN [61] have also performed life cycle analysis showing the energy pay back time for the DSC to be less than one year in south-European climate as compared to over 3 years for silicon solar cells. From these extensive studies, confidence has emerged that the DSCs can match the stability requirements needed to sustain outdoor operation for at least 20 years. This has paved the way for the recent worldwide surge in the industrial development and commercialization of the DSC.

Ongoing and Future Research

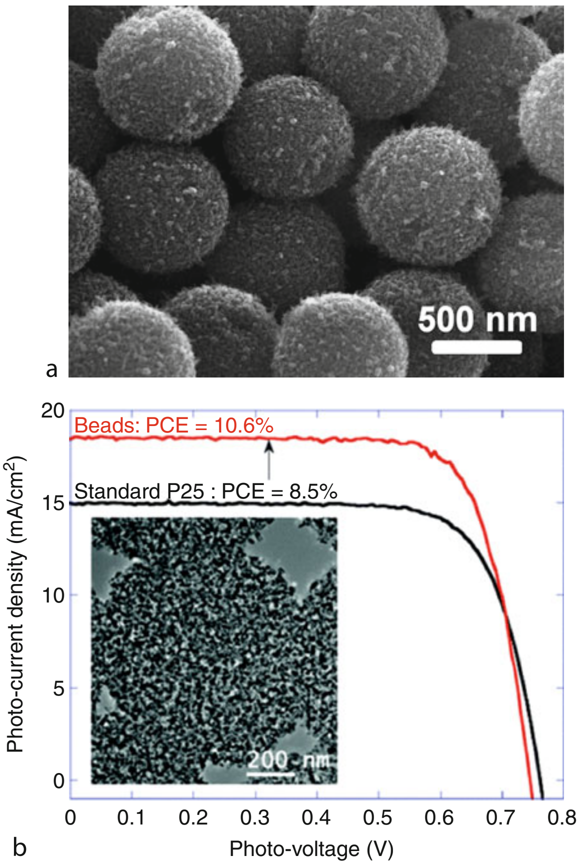
Table 2 lists a number of research topics that are presently under investigation. Below, we illustrate this very active ongoing research with a few examples.

Enhanced light harvesting by advanced nanostructures. Research in this area is presently particularly fertile, and a wealth of new mesoscopic

Mesoscopic Solar Cells. Table 2 A nonexhaustive list of current and future research topics in the field of mesoscopic solar cells

• Enchanced light harvesting by advanced mesostructures
• New sensitizers
• Redox mediators to replace the triidodide/iodide couple
• Alternatives to Pt as electrocatalyst for the counterelectrode
• Solid state sensitized heterojunctions
• Quantum dot injection cells
• Tandem devices
• New solid nanocomposite electrolytes

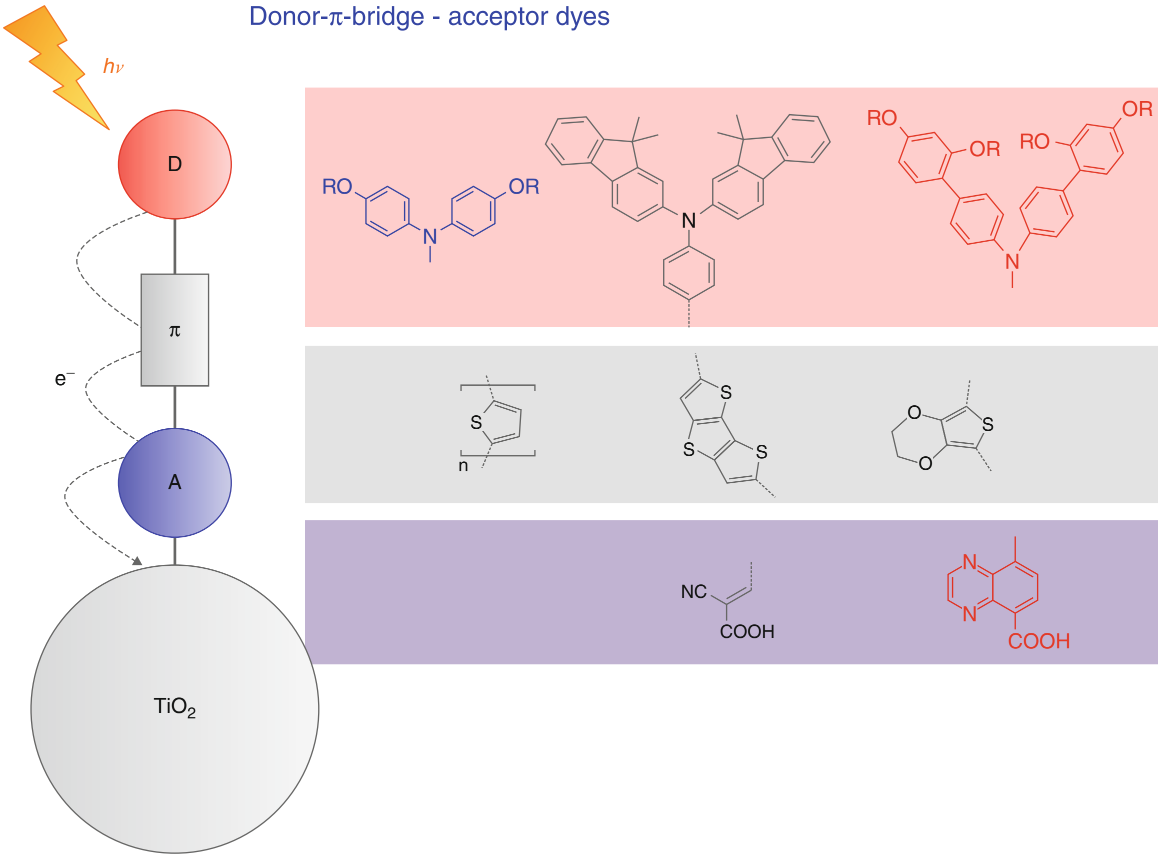
structures ranging from nanorods composed of oxides having a core shell structure to gyroids and nanotubes are presently under active investigation. Here, we restrict ourselves to a discussion of the very intriguing case of mesoporous semiconductor oxide beads. Figure 10a shows an electron microscopy picture of microscopic particles made of TiO_2 and produced by a hydrothermal method using long-chain amine surfactants as templating agents. The burning of the organic component produces mesopores in the 200–1,000-nm-sized beads whose diameter can be varied between 10 and 25 nm [63]. The internal surface area is about 90 m^2/g providing a suitable host for the adsorption of sensitizers. Solar light harvesting by the dye-loaded beads is enhanced by multiple scattering within the assembly of beads [64] leading to an improved response of the photocurrent in particular in the red wavelength region where the light absorption of many commonly used sensitizers is weak. Using a simple film of these anatase beads in conjunction with the C101 sensitizer, strikingly high power conversion efficiencies exceeding 10% have been readily obtained [62]. Figure 10b compares J–V curves for such mesoporous beads and a layer of conventional TiO_2 nanoparticles using the same C101 sensitizer and a similar film thickness. The inset of Fig. 10b shows a transmission electron micrograph illustrating the interconnection of the titania microbeads within the film. Clearly, the mesoporous beads produce higher photocurrents compared to a conventional nanocrystalline TiO_2 film resulting in better conversion efficiencies.



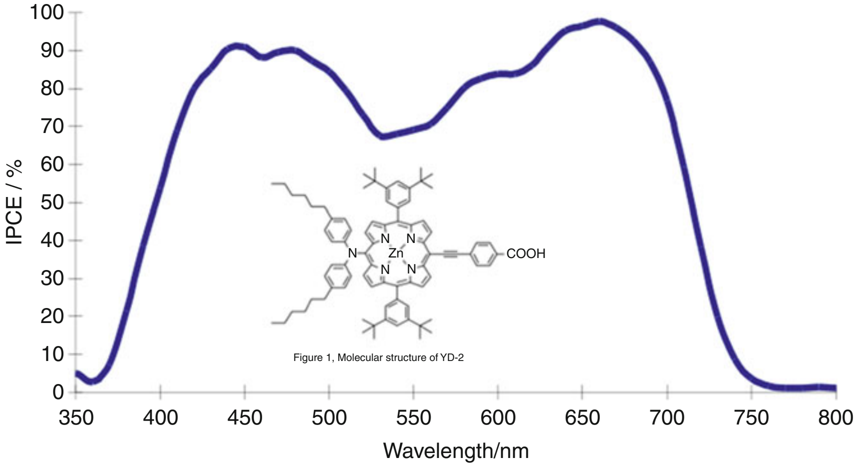
Mesoscopic Solar Cells. Figure 10
(**a**) Scanning electron microscopy picture of an assembly of mesoporous TiO_2 (anatase) beads having an internal surface area of ca. 90 m^2/g. (**b**) Photocurrent voltage curves obtained with similar bead layers using the heteroleptic C101 sensitizer [52]

New sensitizers. Over the last two decades, ruthenium complexes endowed with appropriate ligands and anchoring groups have by far been the preferred choice of charge-transfer sensitizers for mesoscopic solar cells. Recently, however, there has been a surge of interest in organic donor–acceptor dyes [65]. Sensitizers comprising a donor and acceptor group that are bridged by a π-conducting moiety have attracted particular attention. Solar to electric power conversion efficiencies have been sharply increasing, reaching 9.5% in 2008 for the indoline dye D205 [66]. Examples for typical structural elements of such D-p-A dyes are presented in Fig. 11 below. Numerous representatives of this class of compounds have been synthesized [69], and PCE values up to 9.8% have been reported [70].

This rapid development has lead to the discovery of a new class of D-π-A dyes where the π-bridge is constituted by a porphyrin moiety [71]. Power conversion efficiencies of 11% have been reached with the green porphyrin–coded YD-2 whose structure is shown in Fig. 12 along with its IPCE spectrum. Replacing the tert.butyl groups in YD2 by two long chain alkoxy moieties yields the YD2–o-C8 sensitizer reaching a power conversion efficiency of 12.3% as described above in Fig. 9. This is the highest PCE value reported so far rendering this class of D-p-A dyes particularly promising for future applications.



Mesoscopic Solar Cells. Figure 11
Chemical structure for typical molecular components of donor-π-bridge-acceptor (D-π-A) dyes and their linkage to titania nanocrystals through the acceptor group which is coordinatively bound to surface titanium ions



Mesoscopic Solar Cells. Figure 12
Chemical structure and IPCE spectrum of the YD-2 porphyrin, which holds the current power conversion efficiency record of 11% for ruthenium-free sensitizers

The number of suitable options for D-π-A dye structures being very large, state-of-the-art theoretical chemical calculations are being employed as a guide for the selection of the most promising candidates for synthesis. The recently developed scaled opposite-spin configuration interaction singles-doubles technique, abbreviated as SOS-CIS(D) [70], appears to offer great accuracy in calculating the UV–vis spectra of novel D-π-A [71, 72]. This new method will assist the experimentalists in the judicious selection of molecular components to engineer the best-performing push charge-transfer sensitizers.

Summary

The present entry discusses recent research made in molecular photovoltaic cells based on the sensitization of a nanocrystalline wide bandgap semiconductor oxide film by a dye. These cells have now attained efficiencies on the laboratory as well as on the module scale which render them competitive to other thin-film solar cells. Their low cost and ease of production avoiding expensive high vacuum steps should benefit large-scale applications. Impressive stability both under long-term light soaking and high temperature stress has been reached, fostering first industrial applications. These systems will promote the acceptance of renewable energy technologies, not least by setting new standards of convenience and economy.

Acknowledgment

I am grateful to my coworkers and for the support by the organizations listed below. Swiss CTI, CCEM-CH, Swiss National Science Foundation, Swiss Energy Office, US Air Force (European Office of Aerospace Research and Development), FP7 European Joule Program. European Research Council (Advanced Research Grant) GRL Korea (with KRICT) KAUST Center for Advanced Molecular Photovoltaics (CAMP) at Stanford University, Industrial Partners.

Bibliography

1. Grätzel M (2001) Photoelectrochemical cells. Nature 414: 338–344
2. Hagfeldt A, Boschloo G, Sun L, Kloo L, Pettersson H (2010) Dye-sensitized solar cells. Chem Rev 110(11):6595–6663
3. McLeskey JT Jr, Qiao Q (2010) Nanostructured organic solar cells. Nanotechology for Photovoltaics, Loucas Tsakalakos Editor CRC Press 147–185
4. Zhang W, Cheng Y, Yin X, Liu B (2011) Solid-state dye – sensitized solar cells with conjugated polymers as hole-transporting materials. Macromol Chem Phys 212:15–23
5. Mathews N, Varghese B, Sun C, Thavasi V, Andreasson B-P, Sow Ch-H, Ramakrishna S, Mhaisalkar S-G (2010) Oxide nanowire networks and their electronic and optoelectronic characteristics. Nanoscale 2:1984–1998
6. Sekar N, Gehlot VY (2010) Metal complex dyes for dye – sensitized solar cells: recent developments. Resonance 15:819–831
7. Ning Z, Fu Y, Tian H (2010) Improvement of dye – sensitized solar cells: what we know and what we need to know. Energy Environ Sci 3:1170–1181
8. Rowley JG, Farnum BH, Ardo S, Meyer GJ (2010) Iodide chemistry in dye – sensitized solar cells: making and breaking I-I bonds for solar energy conversion. J Phys Chem Lett 1: 3132–3140
9. Halme J, Vahermaa P, Miettunen K, Lund P (2010) Device physics of dye solar cells. Adv Mat 22:E210–E234
10. Wang L, Fang X, Zhang Z (2010) Design methods for large scale dye – sensitized solar modules and the progress of stability research. Renew Sustain En Rev 14:3178–3184
11. Ma BB, Gao R, Wang L-D, Zhu Y-F, Shi Y-T, Geng Y, Dong H-P, Qiu Y (2010) Recent progress in interface modification for dye – sensitized solar cells. Science China: Chem 53:1669–1678
12. Ruehle S, Shalom M, Zaban A (2010) Quantum-dot- sensitized solar cells. Chemphyschem 11:2290–2304
13. Woehrle D, Hild O-R (2010) Energy of the future. Organic solar cells. Chem Unserer Zeit 44:174–189
14. Li X, Wang H, Wu H (2010) Phthalocyanines and their analogs applied in dye – sensitized solar cell. Struct Bond 135:229–274
15. Caramori S, Cristino V, Boaretto R, Argazzi R, Bignozzi C-A, Di Carlo A (2010) New components for dye – sensitized solar cells. Int J Photoenergy 1–17
16. Uzaki K, Nishimura T, Usagawa J, Hayase S, Kono M, Yamaguchi Y (2010) Dye – sensitized solar cells consisting of 3D-electrodes – a review: aiming at high efficiency from the view point of light harvesting and charge collection. J Solar Energy Eng 132:021204
17. Park N-G (2010) Dye – sensitized metal oxide nanostructures and their photoelectrochemical properties. J Korean Electrochem Soc 13:10–18
18. Aguilar RH, Sommeling PM, Kroon JM, Mendes A, Costa CAV (2009) Dye – sensitized solar cells: novel concepts, materials, and state-of-the-art performances. Int J Green Energy 6(3):245–256
19. Desilvestro J (2009) Durability assessment of dye solar cells and modules. In: Miyasaka T (ed) Shin konseputo taiyo denchi to seizo purosesu. Shiemushishuppan, Tokyo, pp 196–205
20. Arakawa H (2009) Weathering resistance of dye – sensitized solar cells. In: Miyasaka T (ed) Shin konseputo taiyo denchi to seizo purosesu. Shiemushishuppan, Tokyo, pp 185–195
21. Yanagida S, Yu Y, Manseki K (2009) Iodine/iodide-free dye-sensitized solar cells. Acc Chem Res 42:1827–1838

22. Grätzel M (2009) Recent advances in sensitized mesoscopic solar cells. Acc Chem Res 42:1788–1798
23. Kalyanasundaram K (2010) Dye-sensitized solar cells. EPFL Press, Lausanne (distributor CRC Press, Boca Raton USA)
24. Desilvestro J, Grätzel M, Kavan L, Moser JE, Augustynski J (1985) Highly efficient sensitization of titanium dioxide. J Am Chem Soc 107:2988–2990
25. Vlachopoulos N, Liska P, Augustynski J, Grätzel M (1988) Very efficient visible light energy harvesting and conversion by spectral sensitization of high surface area polycrystalline titanium dioxide films. J Am Chem Soc 110:1216–1220
26. O'Regan B, Grätzel M (1991) A low-cost, high efficiency solar cell based on dye sensitized colloidal TiO_2 films. Nature 335: 737–740
27. Nazeeruddin MK, Kay A, Rodicio I, Humphrey-Baker R, Müller E, Liska P, Vlachopoulos N, Grätzel M (1993) Conversion of light to electricity by cis-X_2bis(2,2′-bipyridyl-4,4′-dicarboxylate) ruthenium(II) charge transfer sensitizer (X = Cl-, Br-, I-, CN-, and SCN-) on nanocrystalline TiO_2 electrodes. J Am Chem Soc 115:6382–6390
28. Bach U, Lupo D, Comte P, Moser JE, Weissörtel F, Salbeck J, Spreitzert H, Grätzel M (1998) Solid state dye sensitized cell showing high photon to current conversion efficiencies. Nature 395:550
29. Qin P, Linder M, Brinck T, Boschloo G, Hagfeldt A, Sun L (2009) High incident photon-to-current conversion efficiency of p-Type dye-sensitized solar cells based on NiO and organic chromophores. Adv Mat 21:1–4
30. Nusbaumer H, Zakeeruddin SM, Moser J-E, Grätzel M (2003) An alternative efficient redox couple for the dye-sensitized solar cell system. Chem Eur J 9:3756–3763
31. Brugnati M, Caramori S, Cazzanti S, Marchini L, Argazzi R, Bignozzi CA (2007) New components for dye-sensitized solar cells. Int J Photoenergy 2:80756/1–80756/10
32. Feldt SM, Gibson EA, Gabrielsson E, Sun L, Boschloo G, Hagfeldt AJ (2010) Design of organic dyes and cobalt polypyridine redox mediators for high efficiency dye-sensitized solar cells. Am Chem Soc 132:16714–16724
33. Zhang Z, Chen P, Murakami TN, Zakeeruddin SM, Grätzel M (2008) The 2,2,6,6-Tetramethyl-1-piperidinyloxy radical: an efficient, iodine-free redox mediator for dye-sensitized solar cells. Adv Funct Mat 18:341–346
34. Wang M, Chamberland N, Breau L, Moser J-E, Humphry-Baker R, Marsan B, Zakeeruddin S-M, Grätzel M (2010) An organic redox electrolyte to rival triiodide/iodide in dye-sensitized solar cells. Nat Chem 2:385–389
35. Li DM, Li H, Luo YH, Li KX, Meng QB, Armand M, Chen LQ (2010) Non-corrosive, non-absorbing organic redox couple for dye-sensitized solar cells. Adv Funct Mater 20(19):3358
36. Daeneke T, Kwon TH, Holmes AB, Duffy NW, Bach U, Spiccia L (2011) High-efficiency dye-sensitized solar cells with ferrocene-based electrolytes. Nat Chem 3:211–215
37. Nazeeruddin MK, Pechy P, Renouard T, Zakeeruddin SM, Humphry-Baker R, Comte P, Liska P, Cevey L, Costa E, Shklover V, Spiccia L, Deacon GB, Bignozzi CA, Grätzel M (2001) Engineering of efficient panchromatic sensitizers for nanocrystalline TiO2-based solar cells. J Am Chem Soc 123: 1613–1624
38. Rothenberger G, Comte P, Grätzel M (1999) A contribution to the optical design of dye-sensitized nanocrystalline solar cells. Sol En Mat Sol Cells 58:321–336
39. Zukalová M, Procházka J, Zukal A, Yum JH, Kavan L, Grätzel M (2010) Organized mesoporous TiO2 films stabilized by phosphorus: application for dye-sensitized solar cells. J Electrochem Soc 157:H99–H103
40. Galoppini E, Rochford J, Chen H, Saraf G, Lu Y, Hagfeldt A, Boschloo G (2006) Fast electron transport in metal organic vapor deposition grown dye-sensitized ZnO nanorod solar cells. J Phys Chem B 110:16159–16161
41. Shankar K, Bandara J, Paulose M, Wietasch H, Varghese OK, Mor GK, LaTempa TJ, Thelakkat M, Grimes CA (2008) Highly efficient solar cells using TiO_2 nanotube arrays sensitized with a donor-antenna dye. Nano Lett 8:1654–1659
42. Macak JM, Ghicov A, Hahn R, Tsuchiya H, Schmuki P (2006) Photoelectrochemical properties of N-doped self-organized titania nanotube layers with different thicknesses. J Mat Res 21:2824–2828
43. Nelson J, Chandler RE (2004) Random walk models of charge transfer and transport in dye sensitized systems. Coord Chem Rev 248:1181–1194
44. Colodrero S, Mihi A, Haggman L, Ocana M, Boschloo G, Hagfeldt A, Miguez H (2009) Porous one-dimensional photonic crystals improve the power-conversion efficiency of dye-sensitized solar cells. Adv Mat 21:764–770
45. Chiba Y, Islam A, Watanabe Y, Komiya R, Koide N, Han L (2006) Dye sensitized solar cells with conversion efficiency of 11.1%. Jap J Appl Phys Part 2(45):24–28
46. Uchida S (2011) Invited lecture presented at the symposium on nanostructured photosystems at the NTU Singapore symposium on July 26
47. Wang Q, Ito S, Graetzel M, Fabregat-Santiago F, Mora-Sero I, Bisquert J, Bessho T, Imai H (2006) Characteristics of high efficiency dye-sensitized solar cells. J Phys Chem B 110: 25210–25221
48. Han L, Islam A, Chen H, Malapaka C, Chiranjeevi B, Zhang S, Yang X, Yanagida M (2012) High-efficiency dye-sensitized solar cell with a novel co-adsorbent. Energ Environ Sci. Accepted paper. DOI:10.1039/c0xx00000x
49. Yella A, Lee H-W, Tsao HN, Yi C, Chandiran AK, Nazeeruddin MdK, Diau EW-G, Yeh C-Y, Zakeeruddin SM, Grätzel M (2011) Porphyrin-sensitized solar cells with cobalt (II/III)-based redox electrolyte exceed 12 percent efficiency. Science 334:629–634
50. Liska P, Thampi R, Grätzel M, Brémaud D, Rudmann D, Upadhyaya HM, Tiwari AN (2006) Nanocrystalline dye-sensitized solar cell/copper indium gallium selenide thin-film tandem showing greater than 15% conversion efficiency. Appl Phys Lett 88:203103
51. Barber GD, Hoertz PG, Lee S-HA, Abrams NM, Mikulca J, Mallouk TE, Liska P, Zakeeruddin SM, Grätzel M, Ho-Baillie A, Green MA (2011) Utilization of direct and diffuse sunlight in a

dye-sensitized solar cell — silicon photovoltaic hybrid concentrator system. J Phys Chem Lett 2:581–585

52. Ito S, Murakami TN, Comte P, Liska P, Gratzel C, Nazeeruddin MK, Gratzel M (2008) Fabrication of thin film dye sensitized solar cells with solar to electric power conversion efficiency over 10%. Thin Solid Films 516:4613–4619
53. Späth M, Sommeling PM, van Roosmalen JAM et al (2003) Reproducible manufacturing of dye-sensitized solar cells on a semi-automated baseline. Progr Photovoltaics: Res Appl 11:207–220
54. Green MA, Emery K, Hishikawa Y, Warta W (2011) Solar effciency tables (Version 37) Prog. Photovolt: Res. Appl. 19:84–92
55. Lenzmann FO, Kroon JM (2007) Recent advances in dye-sensitized solar cells. Adv Opto-Electr (Recent Advances in Solar Cells) 65073/1–65073/10
56. Grätzel M (2008) Recent applications of nanoscale materials: solar cells. In: Leite RE (ed) Nanostructured materials for electrochemical energy production and storage. Springer, New York, Chapter 1
57. Grätzel M (2006) Photovoltaic performance and long-term stability of dye-sensitized mesocopic solar cells. C. R. Chimie 9:578–583
58. Arakawa H, Yamaguchi T, Okada K, Matsui K, Kitamura T, Tanabe N (2009) Highly durable dye-sensitized solar cells. Fujikura Tech Rev 2009:55–59
59. Harikisun R, Desilvestro H (2011) Long-term stability of dye solar cells. Sol Energ 85:1179–1188
60. http://3gsolar.com/NewsItem.aspx?ID=40
61. De Wild-Scholten MJ, Veltkamp AC (2007) Environmental life cycle analysis of dye sensitized solar devices. www.ecn.nl/publicaties/PdfFetch.aspx?nr=ECN-M–07-081
62. Sauvage F, Chen D, Comte P, Huang F, Heiniger L-P, Cheng Y-B, Caruso RA, Grätzel M (2010) Dye-sensitized solar cells employing a single film of mesoporous TiO_2 beads achieve power conversion efficiencies over 10%. ACS Nano 4(8):4420–4425
63. Chen D, Cao L, Huang F, Imperia P, Cheng Y-B, Caruso RA (2010) Synthesis of monodisperse mesoporous titania beads with controllable diameter, high surface areas, and variable pore diameters (14–23 nm). J Am Chem Soc 132(12):4438–4444
64. Huang F, Chen D, Zhang X-L, Caruso RA, Cheng Y-B (2010) Dual-function scattering layer of submicrometer-sized mesoporous TiO_2 beads for high-efficiency dye-sensitized solar cells. Adv Funct Mat 20(8):1301–1305
65. Qin H, Wenger S, Xu M, Gao F, Jing X, Wang P, Zakeeruddin S-M, Grätzel M (2008) An organic sensitizer with a fused dithienothiophene unit for efficient and stable dye-sensitized solar cells. J Am Chem Soc 130(29):9202–9203
66. Ito S, Miura H, Uchida S, Takata M, Sumioka K, Liska P, Comte P, Pechy P, Gratzel M (2008) High-conversion-efficiency organic dye-sensitized solar cells with a novel indoline dye. Chem Comm 41:5194–5196
67. Yum J-H, Hagberg DP, Moon S-J, Karlsson KM, Marinado T, Sun L, Hagfeldt A, Nazeeruddin MK, Grätzel M (2009) A light-resistant organic sensitizer for solar-cell applications. Angew Chem Int Ed 48:1576–1580
68. Zhang G, Bala H, Cheng Y, Shi D, Lv X, Yu Q, Wang P (2009) High efficiency and stable dye-sensitized solar cells with an organic chromophore featuring a binary π-conjugated spacer. Chem Comm 2198–2200
69. Hsieh C-P, Lu H-P, Chiu C-L, Lee C-W, Chuang S-H, Mai C-L, Yen W-N, Hsu S-J, Diau EW-G, Yeh C-Y (2010) Synthesis and characterization of porphyrin sensitizers. J Mater Chem 20:1127
70. Bessho T, Zakeeruddin SM, Yeh C-Y, Diau EWG, Grätzel M (2010) Highly efficient mesoscopic dye-sensitized solar cells based on donor-acceptor-substituted porphyrins. Angew Chem Int Ed 49:6646–6649
71. Rhee YM, Head-Gordon M (2007) Scaled second-order perturbation corrections to configuration interaction singles: efficient and reliable excitation energy methods. J Phys Chem A 111:5314–5326
72. Casanova D, Rotzinger FP, Grätzel M (2010) Computational study of promising organic dyes for high-performance sensitized solar cells. J Chem Theory Comput 6:1219–1227

Meterology and Wind Power

Erik Lundtang Petersen, Peter Hauge Madsen
Wind Energy Division, Risø DTU National Laboratory
for Sustainable Energy, Technical University of
Denmark, Roskilde, Denmark

Article Outline

Glossary

Atmospheric boundary layer (ABL) Also known as planetary boundary layer (PBL), it is the bottom layer of the atmosphere that is in contact with the surface of the Earth. It extends from 100 m or less in a clear nighttime condition to more the 2 km on a convective sunny day.

Atmospheric surface layer The atmospheric layer closest to the ground up to 50–100 m and where the pressure and Coriolis forces can be neglected in the parameterization of meteorological variables such as the wind profile. The fluxes of the momentum and heat are nearly constant with height.

CFD models Computational Fluid Dynamics models which are mainly used for engineering purposes such as aerodynamic calculations for the flow around a wind turbine blade. They are currently being developed for use for wind studies in very complicated topography.

Climatology The average weather experienced at a place in the course of some chosen run of years.

Coriolis force As air moves from high to low pressure in the northern hemisphere, it is deflected to the right by the Coriolis force. In the southern hemisphere, air moving from high to low pressure is deflected to the left by the Coriolis force. The Coriolis force is caused by the rotation of the Earth.

Downscaling methods The concept of downscaling large-scale analysis and forecasts of weather and climate, such that small-scale features are estimated based on input about large-scale structures of the atmosphere. Two concepts are used: dynamical and statistical.

Dynamic downscaling Use of mesoscale meteorological models to generate high-resolution climate statistics for a specific region and period of time based on, for example, the Global data archive.

Dynamic statistical downscaling Use of mesoscale meteorological models to generate high-resolution climate statistics for a specific region and period of time based on a selected number of weather situations from, for example, the Global data archive.

Geostrophic wind The wind which is in balance between the pressure and the Coriolis forces. It is often close to the wind observed above the PBL by radiosondes and can be calculated from surface pressure measurements.

Global data archive Global or near-global covering climatological and topographic data.

Hub height Height above the ground at the center of the rotor – usually the same as the tower height.

IEC 61400–1 International Standard published by the International Electrotechnical Commission. The Standard specifies essential design requirements to ensure the engineering integrity of wind turbines. Its purpose is to provide an appropriate level of protection against damage from all hazards during the planned lifetime.

Lib files Tables of the two Weibull parameters given for a number of wind direction sectors, heights above ground and terrain roughness classes used in the wind atlas methodology.

Lidar Light detection and ranging. Wind measurement device based on laser – Doppler technology.

Mesoscale model Numerical meteorological models based on the full set of dynamical fluid equations usually covering a region of a few hundred of kilometers and a grid resolution of 2–10 km. An example is the PSU/NCAR mesoscale model (known as MM5) which is a limited-area, nonhydrostatic, terrain-following sigma-coordinate model designed to simulate or predict mesoscale atmospheric circulation.

Microscale model Numerical flow models that can be based on the dynamical fluid equations, for example, CFD models or be based on a linearized version of the fluid equations. An example of a linearized flow model is the BZ (Bessel-zooming-grid) model in WAsP.

Orography The height variations of a terrain.

Power curve Gives the relationship between the net power output of a wind turbine and the wind speed measured at hub height averaged over 10 min.

Reanalysis dataset (Global Data) Time series of the large-scale meteorological situation covering decades. These datasets have been created by assimilating measurement data from around the globe in a dynamical consistent fashion using large-scale numerical models. The primary purpose for the generation of the dataset is to provide a reference for the state of the atmosphere and to identify any features of climate change. For wind energy, the application of the dataset is as a long-term record of large-scale wind conditions.

Reference wind Usually the extreme 10-min average wind speed with a recurrence period of 50 years at turbine hub height. Used in IEC 61400–1 [1] together with the turbulence intensity to define classes for structural loading calculations.

Regional resource assessment Regional resource assessment of wind energy resources means estimating the potential output from a large number of wind turbines distributed over a region. Ideally, this results in detailed, high-resolution, and accurate resource maps, showing the wind resource (yearly and seasonal), the wind resource uncertainty, and areas of enhanced turbulence.

Roughness length The roughness of a terrains commonly parameterized by a length scale called the roughness length. For the logarithmic wind profile, it is the height where the wind is zero.

Siting Siting is a process that includes estimating the mean power produced by specific wind turbines at one or more specific locations. Proper siting of wind turbines with respect to the wind resource requires proper methods for calculating the wind resource, the turbulence conditions, the extreme wind conditions, and the effects of rotor wakes.

The wind atlas method The conventional method used to produce estimates of wind resource on national scales is to analyze wind measurements made at a number of sites around the country as in, for example, the European Wind Atlas [2]. In order for this method to work there needs to be a sufficient quantity of high quality data, covering the country.

Topography The description of shapes and features of the Earth's surface such as orography, land cover, and buildup areas (especially their depiction in maps).

Turbulence The fast variations of the wind vector in all three directions: longitudinal, lateral, and vertical.

Turbulence intensity The ratio between the mean horizontal wind and the standard deviation of the turbulence fluctuations usually measured over a period of 10 min.

WAsP Wind Atlas Analysis and Application Program. Commonly used computer program for siting and regional resource assessment. Developed for the European Wind Atlas.

Weibull probability density function Two-parameter probability density function which very often fits measured wind speed observations well. Is determined by the scale parameter A, which is close to the mean value and the shape parameter k. For $k = 1$ the function is Exponential for $k = 2$ it is the Rayleigh distribution and for $k = 3$ it is close to the Gaussian distribution.

WENE-048 predictions Prediction of the power output from a wind farm hours and days ahead. This term is treated in a separate chapter.

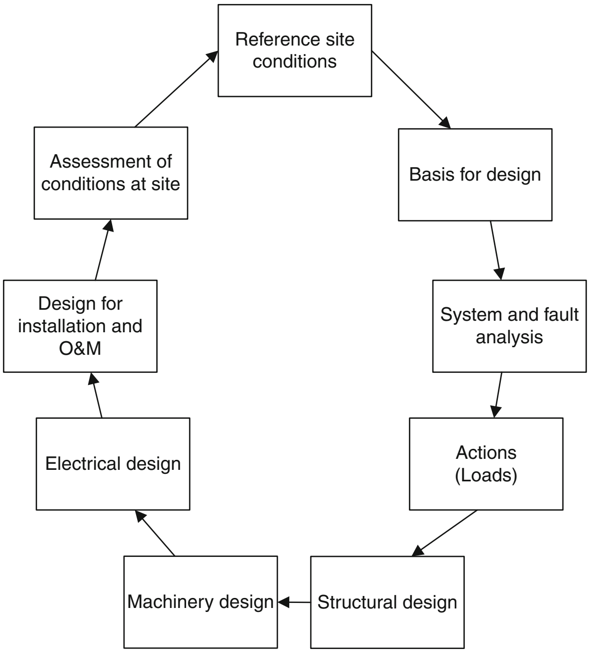
Wind profile The increase of the wind speed (horizontal component of the wind vector) above terrain. In strong wind in an overcast situation (called thermally neutral conditions), it follows a logarithmic law in the lowest 50–100 m. It deviates from the logarithmic when thermal effects become noticeable (stable and unstable conditions). The industry and the IEC 61400–1 often use a power law to describe the height variation. The power 1/7 may be used to represent neutral conditions and a roughness length of 0.03 m.

Definition of the Subject

The utilization of wind energy requires an ability to assess the wind conditions with a high degree of certainty, being paramount for obtaining low risks and high reliability in wind energy project planning. This in turn requires a profound understanding of how atmospheric motions affect the use of wind energy: From the design and operation of the turbines to the spatial integrated renewable energy systems, say, from the dynamic inflow conditions at the turbine rotor to regional resource assessments. The activities necessary for solving the inherent problems in making use of the kinetic energy available in air that passes through the rotor of a large wind turbine during normal operation – or – respectively minimize the problems, has required a huge effort on analytical formulations, experimental activities, and the creation of dedicated numerical tools. Then, through a close cooperation between scientists and engineers, substantial progress has been achieved in the understanding of the interaction between the wind turbines and the atmospheric motions and has brought it to the practical level of international standards, norms, guidelines, and best practices. The huge advances in information sharing technology, in computer power, and in experimental capability have further aided to the fast progress of the discipline which in a broad term can be called: Wind Power Meteorology.

The current fast development in Wind Power Meteorology can be termed: "From Global to Local" which is synonymous with the process that goes from catching the relevant data information from Global dataset, applying the data in meteorological models from global to mesoscale, which in turns produces spatial and temporal information to be used by microscale models to produce the relevant information on the wind conditions for design and for power production at a specific site for specific wind turbines. Hence, this can be described as the two parallel chains of respectively models and in- and output statistics. However, the "model chain": Global – mesoscale – microscale is still in its infancy and fundamental problems have to be sorted out. Therefore, most studies of wind conditions for wind power still rely on the use of local measurements in combination with microscale models. This can ensure that local effects are well described but often have the problem of a too short period of measurements, whereas the model chain often can take advantage of the global datasets which covers decades. They have on the other hand the problem of resolution and the computer time necessary for the models to catch the important wind systems on a sufficient small scale.

This chapter treats the two main wind conditions subjects: wind resources and design conditions. The expected power production from a wind turbine is calculated by means of the measured or modeled probability density function of the wind speed at hub height and the power curve for the wind turbine. Determining the design conditions for the turbine is more involved and the wind parameters of interest such as extreme winds and wind gusts, very fast changes in wind direction, and shears across the rotor of wind speed and wind direction are difficult to measure or model with a sufficient accuracy. A complication is that it is not necessarily the magnitude of these inflow conditions that determines the design of a specific turbine but how they impact the turbine's configuration and operational state, such as start-up events, shutdown events, and its control system in general. A wind turbine is a series-produced industrial product, which cannot be designed according to local conditions in all detail. Instead, wind turbines are designed according to reference conditions. Such reference design requirements for land-based wind turbines are specified in the international design standards given by the International Electrotechnical Commission (IEC) 61400–1 [1]. Here a number of wind classes are specified and the designer has to demonstrate that the design is adequate with respect to the loading of the wind turbine in the various classes. Finally, for the specific wind turbine project it has to be demonstrated that the actual conditions are more benign than the reference design conditions or that the resulting loading does not exceed the strength reserves in the turbine components. Hence the design process of a wind turbine for series-production will start with reference conditions and end with assessment of conformity with actual site conditions, while being more complex as shown in the sketch below:


Reference site conditions
Basis for design
System and fault analysis
Actions (Loads)
Structural design
Machinery design
Electrical design
Design for installation and O&M
Assessment of conditions at site

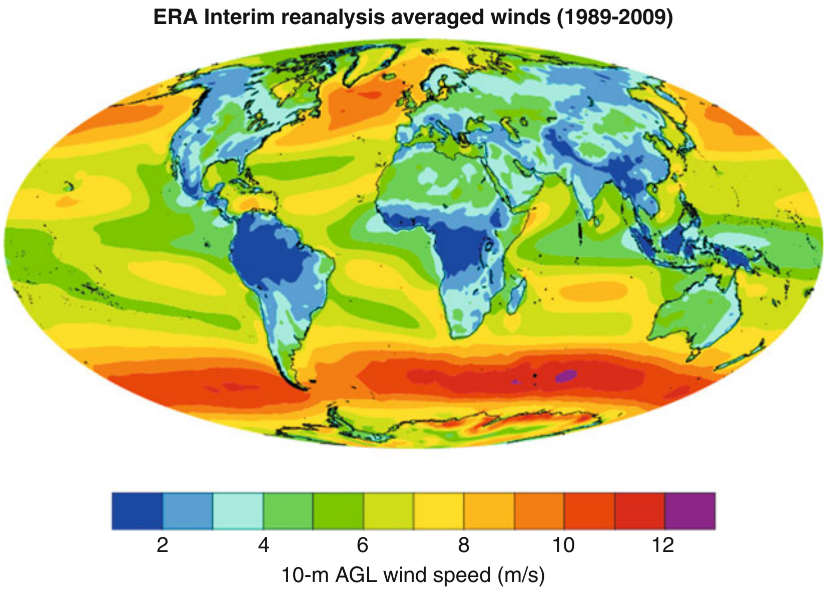
The design aspects of the wind power meteorology are treated in the end of this chapter.

As appetizers, Figs.1 and 2 give, respectively, an overview of the wind resources worldwide and a schematic presentation of many of the concepts that will be treated in this chapter.

Introduction

A recent publication (2009) from the European Wind Energy Association, *The economics of wind energy* [3] states the following:

- *The local wind resource is by far the most important determinant of the profitability of wind energy investments. Just as an oil pump is useless without a sizable oilfield, wind turbines are useless without a powerful*


ERA Interim reanalysis averaged winds (1989-2009)
2
4
6
8
10
12
10-m AGL wind speed (m/s)

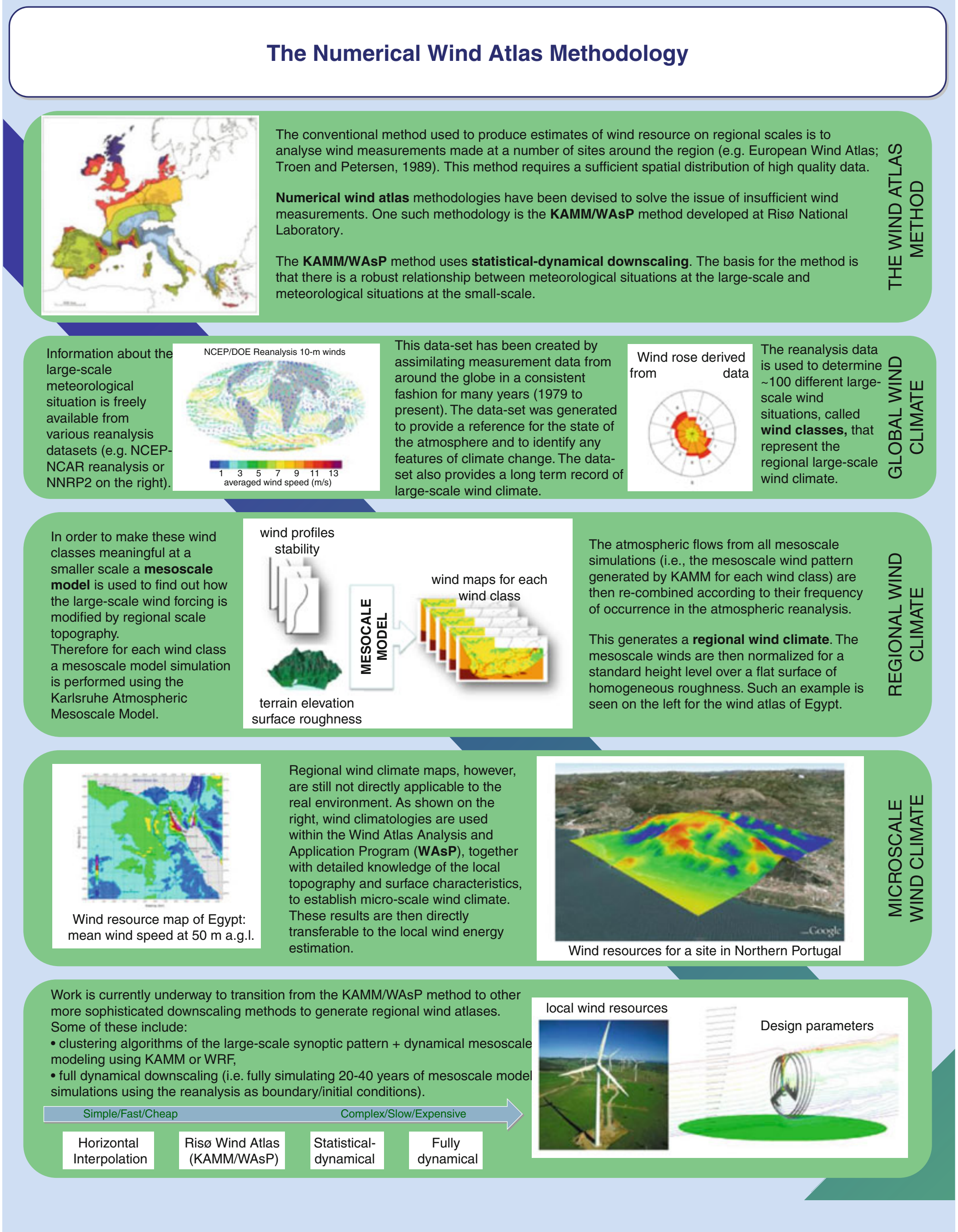
Meterology and Wind Power. Figure 1
The wind climate in terms of the average wind speed at 10 m above ground level was calculated from output from the European Centre for Medium-Range Weather Forecasts (ECMWF; called ERA-Interim) latest reanalysis project. The reanalysis uses a state-of-the-art numerical weather forecast model and techniques that allow taking into account measurement data from around the globe in a consistent fashion from 1989 to the present day. This dataset was generated to provide a reference for the state of the atmosphere and to identify any features of climate change. The dataset is also a long-term record of the large-scale wind climate. Because of the low resolution of the model (1.5° latitude × 1.5° longitude), these data serve only as an indicator for the wind resources of a particular region. To estimate the actual wind energy potential, other higher-resolution models, wind observations, and detailed knowledge of the topography and surface characteristics have to be used. The stronger winds are found between the latitudes of 40° and 50° in both hemispheres; the strongest winds on Earth occur over the southern ocean (sometimes called the "roaring 40s"). Landmasses tend to have weak winds, especially along the equator over the rain forest areas. ECMWF ERA-Interim data used in this project have been obtained from the ECMWF data server (Courtesy Andrea Hahmann)

wind field. The correct micro-siting of each individual wind turbine is therefore crucial for the economics of any wind energy project. In fact, it is beyond dispute that, during the infancy of modern wind industry in 1975–1985, the development of the European Wind Atlas methodology was more important for productivity gains than advances in wind turbine design. The European Wind Atlas method was later formalized in the WAsP computer model for wind resource assessment.

The availability of a suitable wind resource at a turbine site and the ability to accurately calculate it over the 20-year lifespan of the turbine are essential preconditions for the development of wind energy as a major global provider of electricity.

Therefore, the objective for research on wind conditions for wind energy resource and design purposes is to be able to determine the wind conditions with a high accuracy at any place on the Earth which could be a potential site for wind energy exploitation. The wind conditions at a specific site are basically dependent on two factors: the regional climatology and the local topography. Hence, to pursue the objective, one needs climatological and topographical data with global coverage and sufficient resolution in time and space and models which can transform these data into wind energy important key parameters and statistics, which ideally are:

The wind speed probability density functions at relevant heights corresponding to the size of the



The Numerical Wind Atlas Methodology

The conventional method used to produce estimates of wind resource on regional scales is to analyse wind measurements made at a number of sites around the region (e.g. European Wind Atlas; Troen and Petersen, 1989). This method requires a sufficient spatial distribution of high quality data.

Numerical wind atlas methodologies have been devised to solve the issue of insufficient wind measurements. One such methodology is the **KAMM/WAsP** method developed at Risø National Laboratory.

The **KAMM/WAsP** method uses **statistical-dynamical downscaling**. The basis for the method is that there is a robust relationship between meteorological situations at the large-scale and meteorological situations at the small-scale.

Information about the large-scale meteorological situation is freely available from various reanalysis datasets (e.g. NCEP-NCAR reanalysis or NNRP2 on the right).

This data-set has been created by assimilating measurement data from around the globe in a consistent fashion for many years (1979 to present). The data-set was generated to provide a reference for the state of the atmosphere and to identify any features of climate change. The data-set also provides a long term record of large-scale wind climate.

The reanalysis data is used to determine ~100 different large-scale wind situations, called **wind classes,** that represent the regional large-scale wind climate.

In order to make these wind classes meaningful at a smaller scale a **mesoscale model** is used to find out how the large-scale wind forcing is modified by regional scale topography.
Therefore for each wind class a mesoscale model simulation is performed using the Karlsruhe Atmospheric Mesoscale Model.

The atmospheric flows from all mesoscale simulations (i.e., the mesoscale wind pattern generated by KAMM for each wind class) are then re-combined according to their frequency of occurrence in the atmospheric reanalysis.

This generates a **regional wind climate**. The mesoscale winds are then normalized for a standard height level over a flat surface of homogeneous roughness. Such an example is seen on the left for the wind atlas of Egypt.

Wind resource map of Egypt: mean wind speed at 50 m a.g.l.

Regional wind climate maps, however, are still not directly applicable to the real environment. As shown on the right, wind climatologies are used within the Wind Atlas Analysis and Application Program (**WAsP**), together with detailed knowledge of the local topography and surface characteristics, to establish micro-scale wind climate. These results are then directly transferable to the local wind energy estimation.

Wind resources for a site in Northern Portugal

Work is currently underway to transition from the KAMM/WAsP method to other more sophisticated downscaling methods to generate regional wind atlases. Some of these include:
- clustering algorithms of the large-scale synoptic pattern + dynamical mesoscale modeling using KAMM or WRF,
- full dynamical downscaling (i.e. fully simulating 20-40 years of mesoscale model simulations using the reanalysis as boundary/initial conditions).

Meterology and Wind Power. Figure 2
The numerical wind atlas methodology. The *arrow* indicates the process from global data to the estimation of local wind conditions for the wind resource and wind turbine design parameters. The *arrow* also indicates the currently undergoing development from the wind atlas method which is based on the use of measured data on a local or regional scale to the fully dynamical mesoscale-, microscale modeling applying data from The Global Archive to determine local wind conditions. The KAMM model (Karlsruhe atmospheric meteorological model) is developed by University of Karlsruhe (Courtesy Andrea Hahmann)

turbines. Together with the power curve, this is used to calculate the power output of the turbine over its lifetime.

The extreme wind speed statistics which is used to determine the reference wind speed used for the structural design of the wind turbine. Usually, this refers to the wind speed that occurs with a recurrence period of 50 years (the reference wind).

The turbulence conditions, such as the turbulence spectra and turbulence intensity used for the structural and aeroelastic design.

The temporal variation of the wind profile, from ground to top of the rotor: Are there anomalies due to topography (e.g., recirculation) and/or the climate (e.g., stability effects).

Statistics and time series of the temporal variations depicting the daily and seasonal variations.

Regional overview of the wind resource, the turbulence conditions, especially areas with adverse conditions which should be avoided due to possible reductions of the turbines lifetime.

The ability to forecast the power production from a wind power installation hours to days ahead.

In the following, the necessary elements for determining the wind conditions for wind energy purposes will be described: The local meteorological elements, the climatological data, the topographic data, the meteorological models, and the tools that finally enable the users to calculate what is described above under the seven items.

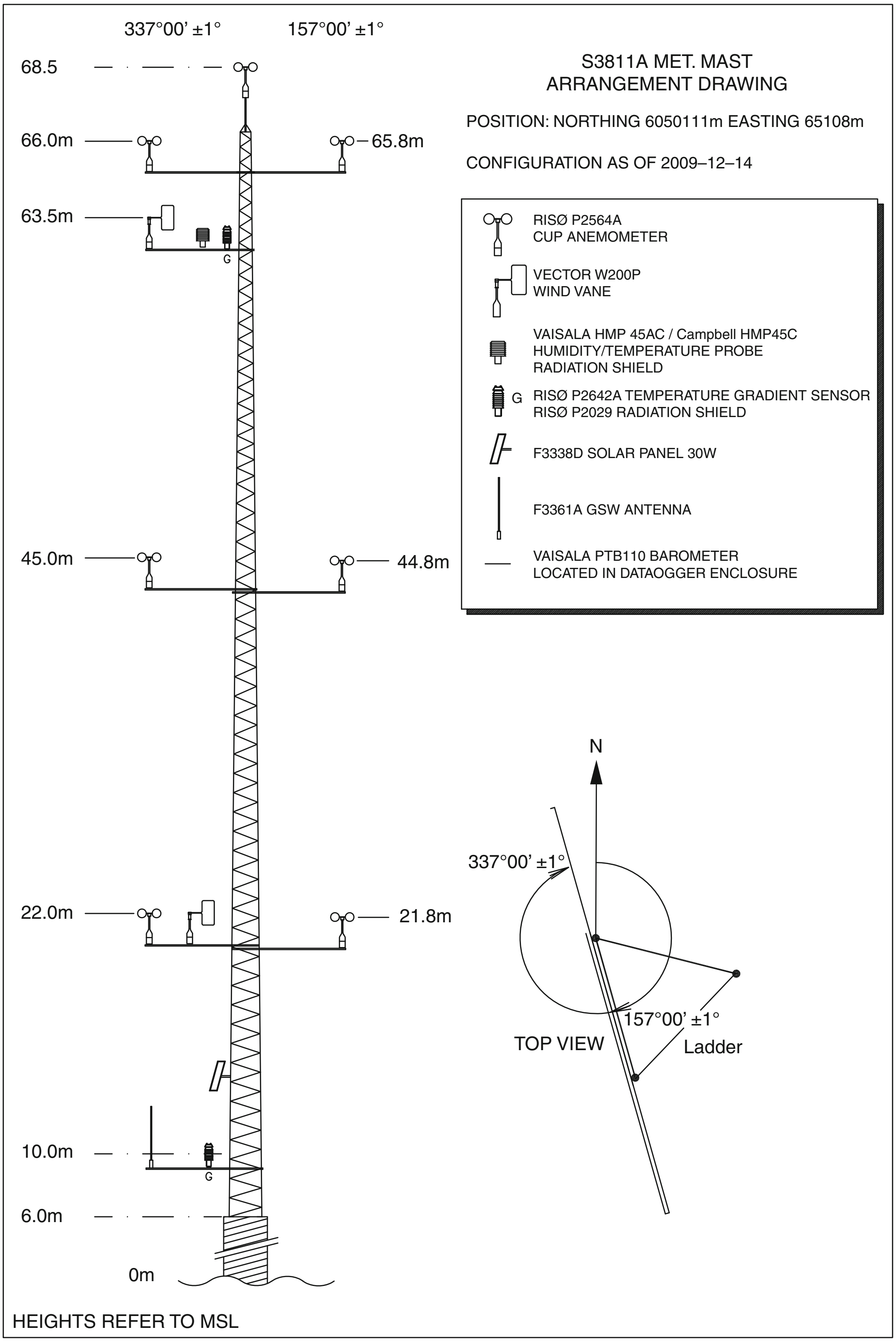
The Local Meteorological Elements

The most important meteorological parameters for the siting of a wind turbine are clearly the wind speed and the wind direction (the wind vector) and their variations with time and height above ground. Other parameters such as temperature, precipitation (rain and snow), and pressure are of importance too – but are less dependent of the specific location. The most direct way of determining the wind conditions at a site is to measure the meteorological parameters for an extended period of time and up to a relevant height, which should be at least to the hub height of the wind turbine (the hub height is the height to the center of the rotor). Figure 3 displays the classical way of measuring the meteorological parameters: A lattice tower with instrument-carrying booms at a number of heights. The length of the measuring period depends on how large uncertainty one will accept in, say, the estimated wind resource. For many wind farm projects, there is only very limited time for measurements before the construction phase; often only 1 year, resulting in a sizable uncertainty. This is mainly caused by the large variation over the year of the wind, but also climatological variations, say, from one decade to another, plays a role and finally, a possible global change in the wind climate may have an effect.

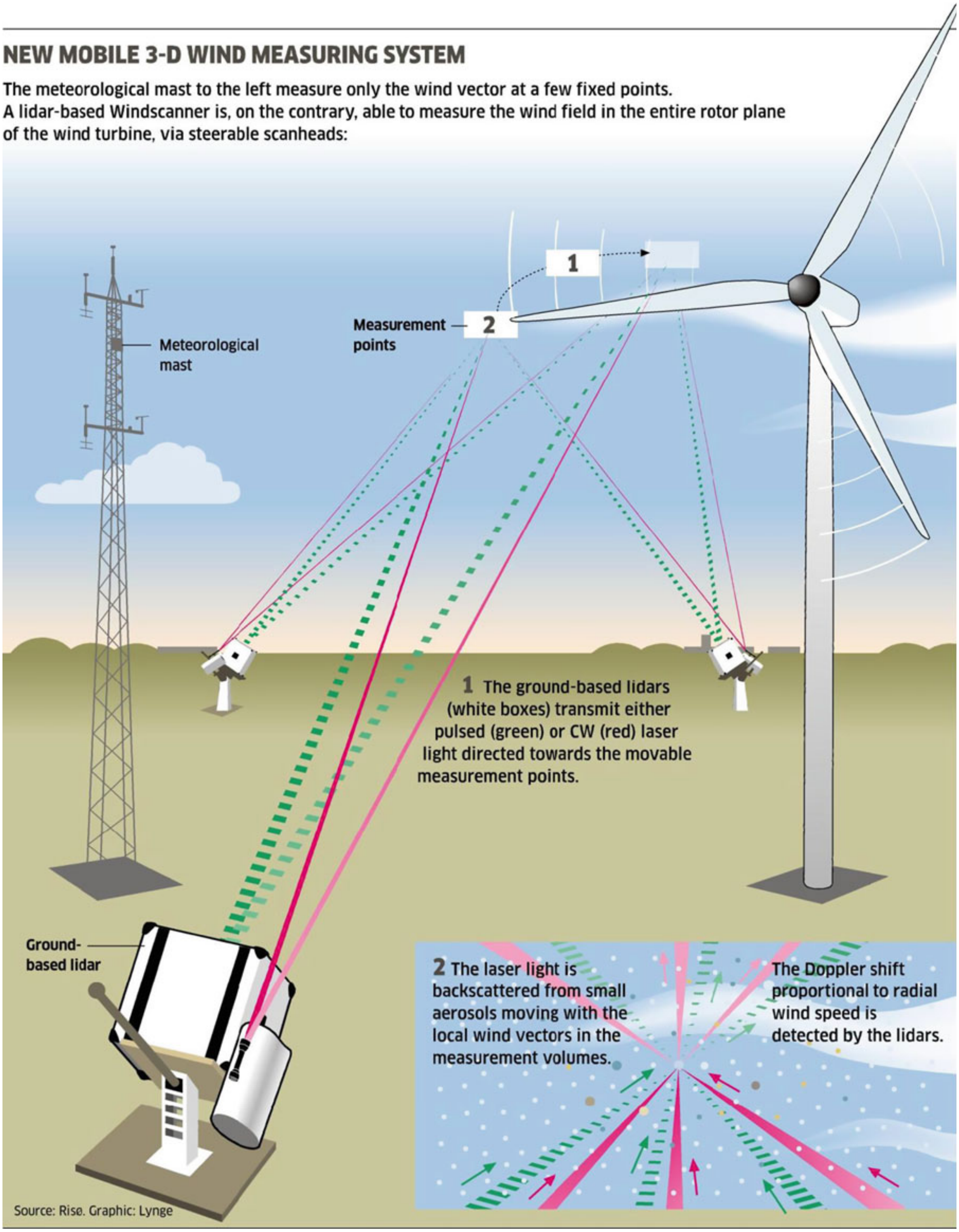
There are several ways to reduce the uncertainty in a short-term measured wind resource. One method uses nearby long-term measured data from one of the climatological stations that are part of the synoptic network together with a microscale program such as the Wind Atlas Analysis and Application Program, which is describe later. Another widely used method is called the Measure – Correlate – Predict method (MCP), where the correlation between the measurements at the mast at the site and the climatological measurements are established for the same period of time and then use to extend the data at the site for the period of the climatological measurements, often 10 years or more. A necessary requirement for this method to work is that the correlations for the most prominent wind directions are very high.

A third way to reduce uncertainty and to extend measured data is to use the meteorological models which are described in the following chapters.

Today's large turbines, where the top of the rotor reaches up to 200 m and future wind turbines which may be even higher, constitute a serious challenge to the ability to determine the wind conditions to such heights. First, the cost of erecting meteorological masts of such a height is often prohibitive and second, because of the necessity then to extrapolate lower measured data to these height one encounters the problem, that the theories which have been used and verified for lower heights, say, up to 100 m, are not valid above this level. This is explained in the next section. Presently, a large effort is therefore directed to develop for wind energy purposes what is termed "Remote sensing" to a sufficient degree of accuracy and robustness. The measuring technique that holds the greatest promises for doing measurements up to several hundred of meters is the Lidar technology



Meterology and Wind Power. Figure 3
Example of an offshore meteorological mast with instrumentation. (*MSL* mean sea level) (Courtesy Anker B. Andersen)



NEW MOBILE 3-D WIND MEASURING SYSTEM

The meteorological mast to the left measure only the wind vector at a few fixed points.
A lidar-based Windscanner is, on the contrary, able to measure the wind field in the entire rotor plane of the wind turbine, via steerable scanheads:

1 The ground-based lidars (white boxes) transmit either pulsed (green) or CW (red) laser light directed towards the movable measurement points.

2 The laser light is backscattered from small aerosols moving with the local wind vectors in the measurement volumes.

The Doppler shift proportional to radial wind speed is detected by the lidars.

Source: Risø. Graphic: Lynge

Meterology and Wind Power. Figure 4
The principle for using three lidars (WindScanner) to detect the three-dimensional wind field around a wind turbine (Courtesy Torben Mikkelsen)

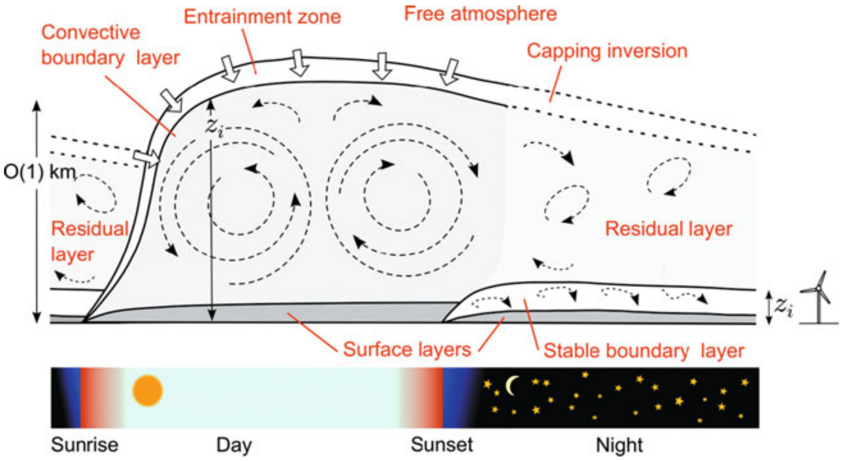
(Light Detection and Ranging). The principle is simple: Emitted laser light is reflected by particles in the atmosphere and due to the Doppler Effect, one can deduce the velocity of the particles by the change in the frequency of the reflected light. For a single lidar pointing vertically the technique is "range gating," this is to say that one can look at particles in several layers above ground and thereby establish the vertical variation of the wind speed with height, that is, the wind profile. This technique has been extended such that the lidar measures in many directions and heights almost instantaneously. When three of these instruments are combined, then the three-dimensional wind vector can be measured over the rotor plane and thus can give information for both the resource calculations and the design requirement. The technique is depicted in Fig. 4.

Figure 5 gives a hint to why the height of the large turbines is a problem for the determination of the wind conditions which is actually in the core of the discipline "Boundary Layer Meteology" which will only be touched briefly here. The reader is referred to the references [4] and [5].

What is depicted in the figure is the diurnal development of the atmospheric boundary layer (ABL) on a clear day over a simple plain topography, for example, an agricultural field with few height variations. This layer is defined as the layer where the underlying surface has a direct influence on the wind and turbulence and other meteorological parameters. Above this layer, the wind is nearly unaffected by the local surface conditions and here blows the free wind, the so-called geostrophic wind.

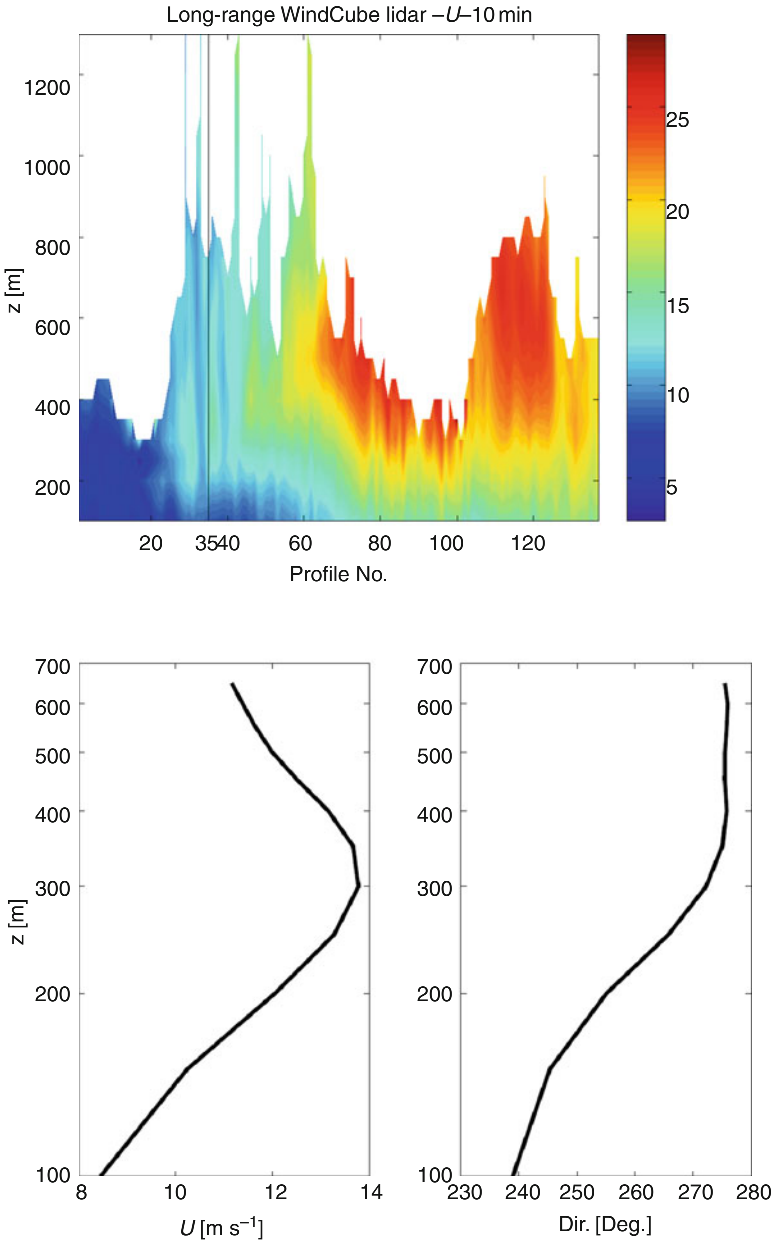
The geostrophic wind is the theoretical wind that would result from an exact balance between the Coriolis force and the pressure gradient force. This condition is called geostrophic balance. The geostrophic wind is directed parallel to isobars (lines of constant pressure at a given height). This balance seldom holds exactly in nature. The true wind almost always differs from the geostrophic wind due to other forces such as friction from the ground. Thus, the actual wind would equal the geostrophic wind only if there were no friction and the isobars were perfectly straight. Despite this, much of the atmosphere outside the tropics is close to geostrophic flow much of the time and it is a valuable first approximation.

The thickness of the atmospheric boundary layer has a distinct diurnal variation, but is also very dependent on the present weather system: On a sunny day with low wind it can be more than 2 km thick, on a day with high winds it will be around 1 km thick, and during a calm night it can be 100 m or less. The conditions are referred to as convective, neutral, and stable conditions. Close to the ground is the surface layer. Here the wind is dominated by the presence of the ground, that is, the roughness and the temperature of the surface and the resulting heatflux. The fast temporal variations of the wind velocity, called the



Meterology and Wind Power. Figure 5
The diurnal variation of the planetary boundary layer over not too complicated topography and local climatology (Courtesy Morten Nielsen)

turbulence, transports energy and hence determines the variation of the wind speed with height – the wind shear – which on the other hand is the mechanism for creating turbulence, the so-called mechanically produced turbulence. In this layer, the wind is considered not to be influenced by the Coriolis Effect and the overall pressure gradient. With the wind speed being in the interval of interest for a wind turbine, that is, between 5 and 25 m/s, this layer can be up to 100 m high and the wind profile is logarithmic with height, for example, the variation from 1 to 10 m height is the same as the variation from 10 to 100 m; see Fig. 6.

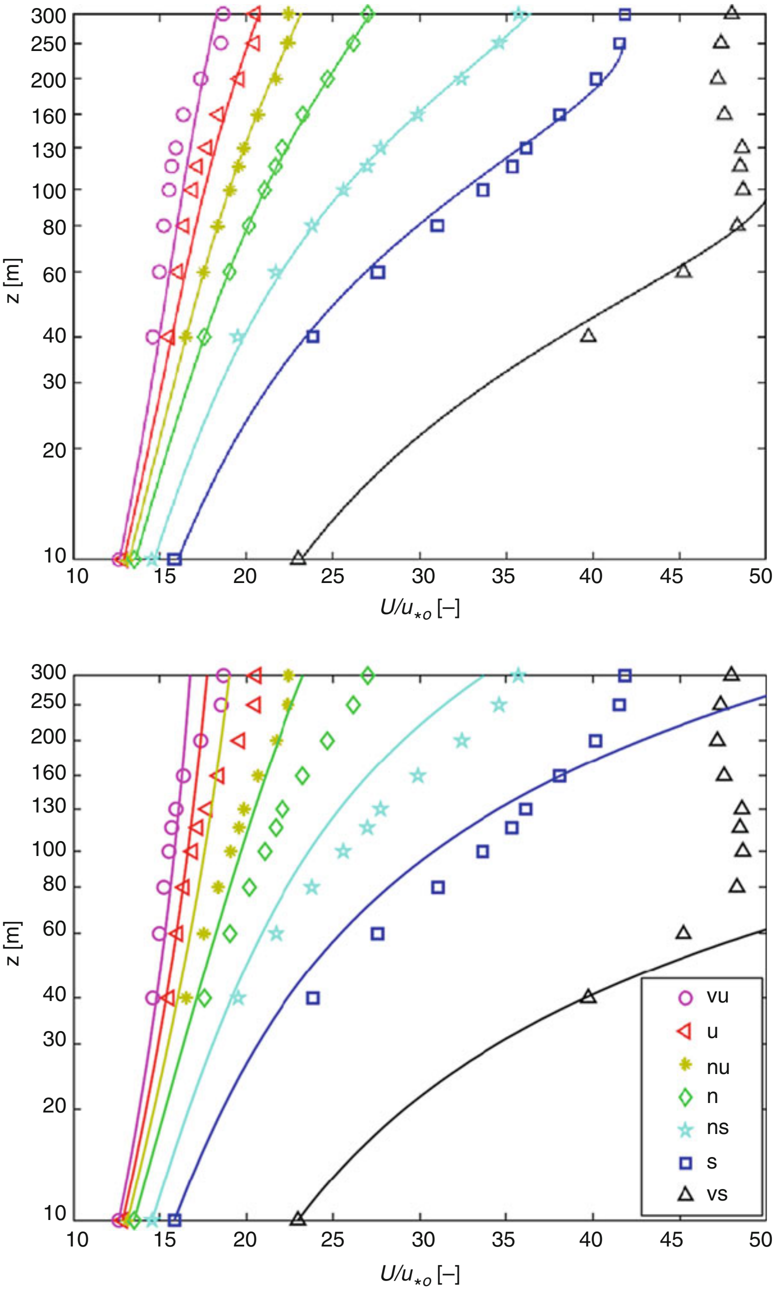


Meterology and Wind Power. Figure 6
Lider measurements (*upper part*) and the derived wind speed and wind direction (*lower part*). The profile is indicated as scan no. 35 (Courtesy Alfredo Peña Diaz)

This is true if the variation of the air temperature is not influencing the production of turbulence, that is, the surface layer is said to be neutral and a parcel of air if moved will stay in equilibrium with the new surroundings. If, on the other hand the temperature decreases fast with height or increases with height there are unstable and stable conditions, respectively, and the turbulence is said to be influenced by thermal effects which in turn deflects the wind profile from the logarithmic height variation. This is well understood and there is a well-founded theory for it called the Monin–Obukov similarity theory [6]. But – and here the problem comes – it only explains what goes on in the lowest 50–100 m. When one moves further up, the wind is also influenced by the Coriolis Effect and the distance to the top of the atmospheric boundary layer. Furthermore, it might be influenced by the variation of the temperature with height above the boundary layer through the processes in the entrainment zone and by creation of internal waves from the action of turbulence on the free atmosphere. Hence the modeling of the wind profile above the surface layer is today far from an established procedure even under well-defined topographic and climatological conditions. For more complicated topography and climatology the knowledge is even less. And the wind conditions over the oceans are turning up to be much more complicated than thought of when the first offshore wind farms were built. Doing meteorological measurements at sea is much more costly and difficult than over land but they are necessary if one shall be able to establish the wind conditions at sea with low uncertainty. For both land and sea the lidar technique holds great promises, but it has it shortcomings: The further away from the instrument the wind is measured, the larger is the volume of air with particles that contributes to the measurement, and therefore this wind is an average over this volume. Figure 6, gives an example of measured wind profiles during a full day up to several hundred meters.

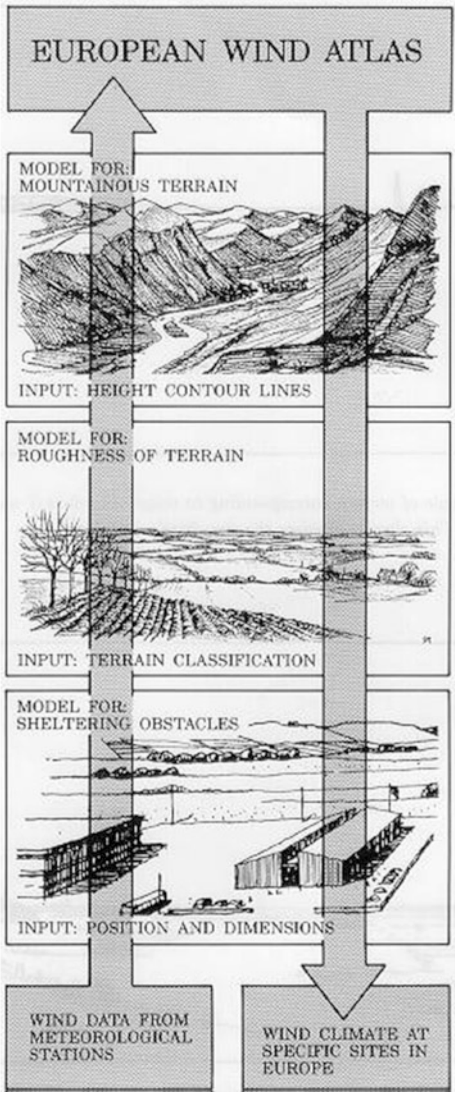
Figure 7 shows examples of measured wind profiles at a Danish North Sea coast together with modeled profiles, where the model takes into account the thermal structure and the height of the atmospheric boundary layer. The boundary layer height is a difficult parameter to determine not to say to get its climatological statistics. For the measured wind speed profiles, the boundary layer height has been determined as the one that gives the best profile fits. For in situ determination of the boundary layer height, remote sensing instruments are required such as lidars, sodars (instrument that emits a sound pulse and listens to the sound reflected from temperature variations with height above ground), and ceilometers, which is a laser-based instrument, basically used to measure the base of clouds. All this is very cumbersome and climatological statistics of the boundary layer height for selected wind farm sites do not exist. Hopefully, satellite measuring techniques will be developed to help out on this.

Now, imagine an ideal sample of wind climatology measured to larger heights, say 200 m, and over a period of 10 years. Definitively, one will be able to perform many of the necessary calculations to determine the wind resource and the wind design conditions at the specific site of the meteorological mast. But how about the conditions at the wind turbines which are to be erected in the large wind farm around the mast? The furthest wind turbine can easily be 10 km away from the mast. If the climatology of the larger scale wind can be considered to be uniform over the wind farm area and if the topography only has small height variations and the roughness is uniform, then it is permissible to use the mast statistics for the whole wind farm. But this is seldom the case; often the topography has variations in orography and land cover such that it is necessary to take this into account when applying the mast statistics to various locations in the wind farm. Even for relatively small wind farms in hilly or mountainous terrain, the variation in the produced power can be more than 25% between the turbines. Therefore, one needs models that can transform the measured wind statistics from the mast location to the locations of all the turbines in the wind farm. Usually this is both a horizontal and a vertical transformation because for economical reasons (and due to some national recommendations) the height of the mast is often chosen to be two-thirds of the hub height. A methodology to perform this task was developed during the creation of the European Wind Atlas (1982–1989) [2] and was made publicly available in 1987 through the PC program WAsP. The basic idea behind the "wind atlas method" is that the wind climatology at a site is "cleaned" from the influence of the surrounding topography to give a generalized regional wind climatology which then could be used at another site, say, tens of kilometers



Meterology and Wind Power. Figure 7

Wind profiles measured at the Danish wind turbine test site Høvsøre by means of anemometers up to 165 m and lidars up to 300 m. Both figures show measurements and theoretical derived profiles. At the *top* figure, the height of the boundary layer *Zi* is an important parameter, whereas at the *bottom* figure the conventional theory (Monin–Obhkov) which is only valid up to 50–100 m is applied. Clearly, *Zi* is an important parameter. Stability explanation: *vu* very unstable, *u* unstable, *nu* neutral to the unstable side, *n* neutral, *ns* neutral to the stable side, *s* stable, and *vs* very unstable (Courtesy Sven-Erik Gryning) away. Then for the new site, the effect of topography at that site is introduced. The procedure is given in Fig. 8 and a short description of its use for the European Wind Atlas is in the next paragraph.


EUROPEAN WIND ATLAS
MODEL FOR: MOUNTAINOUS TERRAIN
INPUT: HEIGHT CONTOUR LINES
MODEL FOR: ROUGHNESS OF TERRAIN
INPUT: TERRAIN CLASSIFICATION
MODEL FOR: SHELTERING OBSTACLES
INPUT: POSITION AND DIMENSIONS
WIND DATA FROM METEOROLOGICAL STATIONS
WIND CLIMATE AT SPECIFIC SITES IN EUROPE

Meterology and Wind Power. Figure 8
A schematic presentation of the wind atlas method used for the transformation of meteorological data. The *left* arrow illustrates the process of "cleaning" the wind data for local conditions. The inverse calculation is illustrated by the *right-hand arrow*: The "cleaned" regionally representative data are used for the calculation of wind conditions at a specific location by introducing the new local conditions

Modeling the Wind Conditions

The state of the atmosphere is well described by seven variables: pressure, temperature, density, water vapor content, two horizontal velocity components, and the vertical velocity component; all functions of time and position. The behavior of these seven variables is governed by seven equations: the equation of state, the first law of thermodynamics, three components of Newton's second law, and the continuity equations for mass and water substance. These equations are mathematical relations between each atmospheric variable and their temporal and spatial derivatives. Mathematical models of the atmosphere can be obtained by integrating the relevant equations with special initial and boundary conditions. The equations can be solved numerically by forward marching in time, using the time rates of change of the variables; the derivatives are replaced by numerical approximations, and changes of the variables over a certain time interval are computed repeatedly as long as needed.

The atmosphere contains motions with scales varying from about 1 mm to thousands of kilometers. Ideally, mathematical models should be constructed from observations with 1 mm spatial and with a fraction of a second temporal resolution. Clearly, this is impossible in practice, and models are constructed separately for systems on different scales. Thus there are models for covering the whole Globe, regional models covering continents, mesoscale models with scales from one to several hundred kilometers and finally microscale models that can resolve topographic and flow features down to a few meters. This is illustrated in Fig. 2 and described in [7].

To solve the differential equations one need a set of initial and boundary conditions. The global models which are run continuously by the large meteorological forecasting centers such as the European Centre for Medium-Range Weather Forecasts (ECMWF) use as initial conditions the millions of measurements from all over the globe shared via the World Meteorological Organization (WMO). There is a great variety of measuring devises covering satellites, ships, radars, buoys, conventional anemometers on masts, etc.; the satellites provide the major part of the data; in all 10^{12} data are collected every hour.

The boundary conditions are provided by global databases of topography (Table 2) and for the oceans, by the sea surface temperature.

In a global circulation model (GCM), the evolution of pressure, wind, temperature, humidity, clouds, rain, and other elements of the weather is computed step by step in a grid with the resolution of approximately 100 km, so as to form a complete, three-dimensional picture of the atmospheric state, evolving with time according to the laws of physics. Such models are used to predict the weather over some days. Another application developed over the last decades is to use such models on a data assimilation setup application of particular interest here.

Model trajectories can be treated in the same way as observations. In particular, averages and other statistical characteristics can be computed for any model variable or combination of variables anywhere within the domain of the model. Similar completeness and coverage cannot be provided by an observation system.

The denser the grid, the smaller is the domain represented by each grid point, and the finer, in general, is the scale of atmospheric motion systems and structures that can be simulated. Still, the model output is not strictly comparable to observations, because the latter will always be influenced by structures and processes beyond the reach of the model – dependent on the distance between grid points. Therefore, in order to resolve atmospheric motions on smaller and smaller scales, it is necessary to operate with a chain of models. In addition, the density in time and space of the observations has a profound influence on the general resolution of the model output. The spatial resolution of topography, that is, land cover and orography also influences what can be resolved by the models.

Still there will always exist processes which have a smaller extent than the size of the computational grid but still are so important for the physics that they have to be taking into account in the modeling process. This is done by the so-called sub-grid parameterization. For example, processes such as convective rain showers of the size of a few kilometers will have to be parameterized in the global models and mesoscale models with a low resolution.

The next step in the model chain is to use the model output from a global model as initial conditions for a regional/mesoscale model. These steps and the resulting higher resolution in the model output are illustrated in Fig. 2. The final step – crucial for wind energy applications is for the mesoscale models to provide the input to a microscale model. Therefore, the core of the downscaling techniques is twofold: First, in the best way to represent the output from the mesoscale model in statistical terms (taking into account thermal stratification, boundary layer structure, and topography) and second to be linked to a microscale model that is capable of using the mesoscale statistics and provide important local siting parameters, such as average and seasonal resource statistics, information on daily and seasonal variability, turbulence and extreme value statistics. An important aspect of the modeling is the validation of the local siting parameters with measurements (if available) and the estimation of uncertainty in the model chain and in the final results.

The output from the mesoscale models is a four-dimensional space – time matrix representing the state of the atmosphere. These data include wind fields at several heights, boundary layer structure parameters, surfaces fluxes, etc. In parallel, the microscale model makes use of the full set of statistics from the mesoscale model and handles effects due to the much more complex topography which generally exist on a scale less than that resolved by the mesoscale model, in order to provide wind characteristics for siting with respect to resource and loads (turbulence, extreme wind, and wind shear). The main problem in the interface between mesoscale and microscale models is to determine the optimal statistical representation of the time series data and important temporal variations, such as seasonal and daily patterns, all dependent on local conditions. Another problem is the ability of the mesoscale model to resolve physical processes on a scale of the order of the numerical grid interval – such as thermally induced flows. Due to the numerical schemes which are used to keep numerical instabilities down, the "true" resolution of a mesoscale model is seldom better than five to seven times the grid length. A typical grid length is 5 km, hence physical processes such as local winds in mountains cannot be resolved and recourse has to be taken to traditional measurements.

The output from the microscale model is validated using available data whenever possible, including data from high masts especially in areas of complex terrain and offshore.

What has been described up to now is the so-called Dynamic Downscaling Method and both this method and the Statistical Downscaling Method are depicted in

Fig. 2. Presently, the latter method is used in most applications, mainly due to a lesser demand on computer power, but the dynamic method is being extensively explored and developed for wind energy purposes due to the much more complete set of wind conditions it can provide.

A large number of mesoscale models are in use, but the most widely used is the WRF (Weather Research Forecast) model developed by the National Center for Atmospheric Research in Boulder, Colorado. It is an open source model and extensive courses are given on its use because it requires highly skilled professionals to run any mesoscale model.

Regional Wind Resource Mapping by Means of the Model Chain

In this paragraph, as a text book example, the concept of The Model Chain is demonstrated by its application in a recent (2010) and very comprehensive study headed by the Finnish Meteorological Institute leading to The Finnish Wind Atlas, illustrated in Fig. 9.

Three different weather prediction models and a microscale model have been used in the production of the Finnish Wind Atlas:

- A Global Model: IFS (Integrated Forecast System) of the European Centre for Medium-Range Weather Forecasts (ECMWF)
- A Regional Model: HIRLAM (High Resolution Limited Area Model of a number of European meteorological institutes, among those FMI: Finnish meteorological Institute)
- A Mesoscale Model: AROME (Applications of Research to Operations at MEsoscale of a number of meteorological institutes and research laboratories, among those, Meteo France)
- A Microscale Model: WAsP (Wind Atlas Analysis and Application Programme of Risø DTU)

The numerical weather prediction models, HIRLAM and AROME, are dynamical forecast models, in which the atmospheric physics is described as accurately as possible for weather forecasting purpose. The WAsP model modifies the wind time series produced by AROME in order to generate the climatological description of wind conditions. WAsP takes into account the effects due to local terrain roughness and topography.

The base material contains 48 months of weather model simulations, in which the winds and other variables are calculated in 2.5 × 2.5 km^2 grid boxes. Furthermore, the results of the weather model are post-processed in 250 × 250 m^2 grid boxes in predefined areas by using WAsP model.

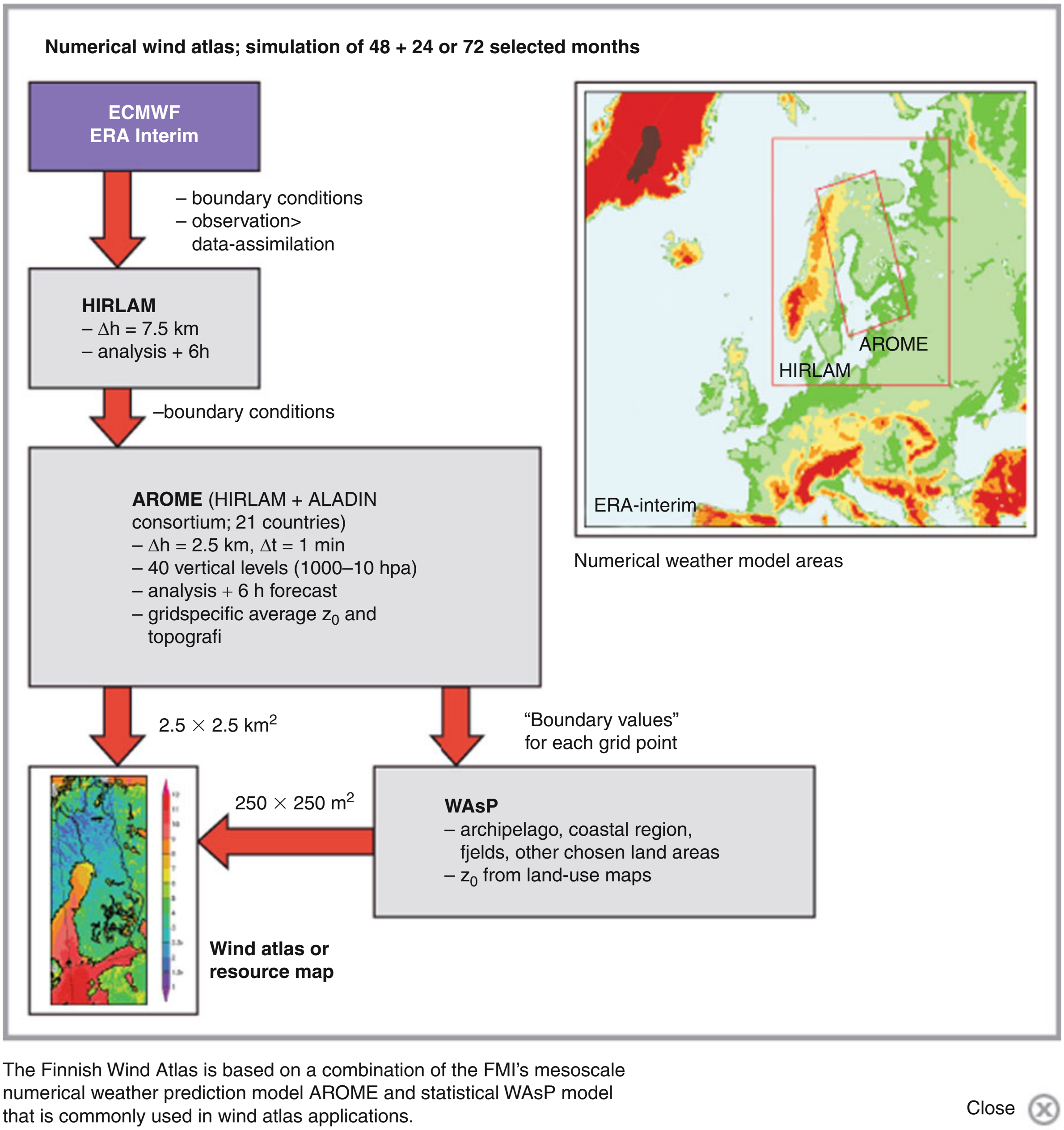
All the wind atlas periods were simulated by AROME in 6-h sequential pieces, where atmospheric state is stored after 3 and 6 h of simulation. The initial state of each 6-h subperiod and the required boundary conditions were produced by using weather observations. FMI's weather prediction model HIRLAM was used in this analysis process. The computational domain of HIRLAM covered the Northern Europe with 7.5-km grid spacing. The data required at the edge of the HIRLAM model was taken from ERA-Interim reanalysis that is produced by European Centre for Medium-Range Weather Forecasts (ECMWF). ERA-Interim data covers the whole globe with 80-km grid spacing.

The Web-based Finnish Wind Atlas includes monthly wind statistics and estimates for wind energy production (MWh) for three different sizes of wind turbine (capacities of 1, 3, and 5 MW). These values are given in seven height levels, 50, 75, 100, 125, 150, 200, and 400 m, and in total of 63,550 points each representing 2.5 × 2.5 km^2 area. Furthermore, the mean wind speed is given also in 250 × 250 m^2 grid in archipelago, field, and coastal regions and in the chosen land regions. These values are processed from AROME-produced wind information by using mesoscale Lib-software and the WAsP model.

Wind information seen in the wind atlas represents mean monthly and annual wind conditions over last 50 years. Wind conditions of an individual year may deviate considerably from the mean conditions. Furthermore, the actual wind conditions may differ from the mean conditions presented in wind atlas during the lifetime of a windmill. The possible effect of climate change is not taken into account in the values of the wind atlas.

Regional Wind Resource Mapping by means of the Climatological Measurements

As described above mesoscale models may not be applicable in certain topographical and climatological settings due to the grid resolution. Then resource has to be taken to use local measurements (provided they



Numerical wind atlas; simulation of 48 + 24 or 72 selected months

ECMWF
ERA Interim

– boundary conditions
– observation>
data-assimilation

HIRLAM
– Δh = 7.5 km
– analysis + 6h

–boundary conditions

AROME (HIRLAM + ALADIN consortium; 21 countries)
– Δh = 2.5 km, Δt = 1 min
– 40 vertical levels (1000–10 hpa)
– analysis + 6 h forecast
– gridspecific average z_0 and topografi

2.5×2.5 km^2

“Boundary values”
for each grid point

WAsP
– archipelago, coastal region, fjelds, other chosen land areas
– z_0 from land-use maps

250×250 m^2

Wind atlas or resource map

Numerical weather model areas

The Finnish Wind Atlas is based on a combination of the FMI’s mesoscale numerical weather prediction model AROME and statistical WAsP model that is commonly used in wind atlas applications.

Meterology and Wind Power. Figure 9
The Finnish Wind Atlas showing the way from the global data – ERA Interim – to the microscale model result with a resolution of 250 m. The atlas is based on the combination of FMI’s mesoscale numerical weather prediction model AROME and the microscale model WAsP. Finnish Meteorological Institution (FMI) (Courtesy Bengt Tammelin)

exist). During the international project, *The European Wind Atlas* with the aim of establishing the meteorological basis for the assessment of the wind resources of the European Union (1981–1989) a methodology – called the *wind atlas method* – was developed resulting in a comprehensive set of models for the horizontal and vertical extrapolation of meteorological data and estimation of wind resources.

The models are based on the physical principles for flow in the atmospheric boundary layer and they take into account the effect of different surface conditions, sheltering effects due to buildings and other obstacles, and the modification of the wind imposed by the specific variations of height of the ground around the meteorological station in question. Figure 8 illustrates the application of the wind atlas method, following a procedure in which the regional wind climatologies are used as input to the models to produce site-specific wind climatologies. The models and the described methodology constitute the microscale program WAsP.

The European Wind Atlas [2] covers land area of about 2.25 million km^2. It employed surface observations of wind speed and direction, measured over a 10-year period, to determine the wind climate at about 190 European meteorological stations. Then using the wind atlas method, the wind climates were subsequently referenced to a common set of standard topographical conditions, that is, they were expressed as Weibull *A*-and *k*-parameters for five heights and 12 30° sectors over four different values of surface roughness. Wind resource estimates for other sites can then be obtained invoking the same method to introduce the site-specific topography of these sites. Figure 10 shows the overview map from the European Wind Atlas.

The publication of the overview map and the table had a profound influence on European decision makers by showing that it is possible to find location with good wind resources almost everywhere if the right topographical settings are selected. This knowledge has an important implication for any modeling of wind resources, such as mesoscale modeling: The coarser the resolution (large grid cells) the meteorological models work with, the smaller becomes the average wind resource inside the grid cell and this is because that favorable locations such as small hills are smoothed out.

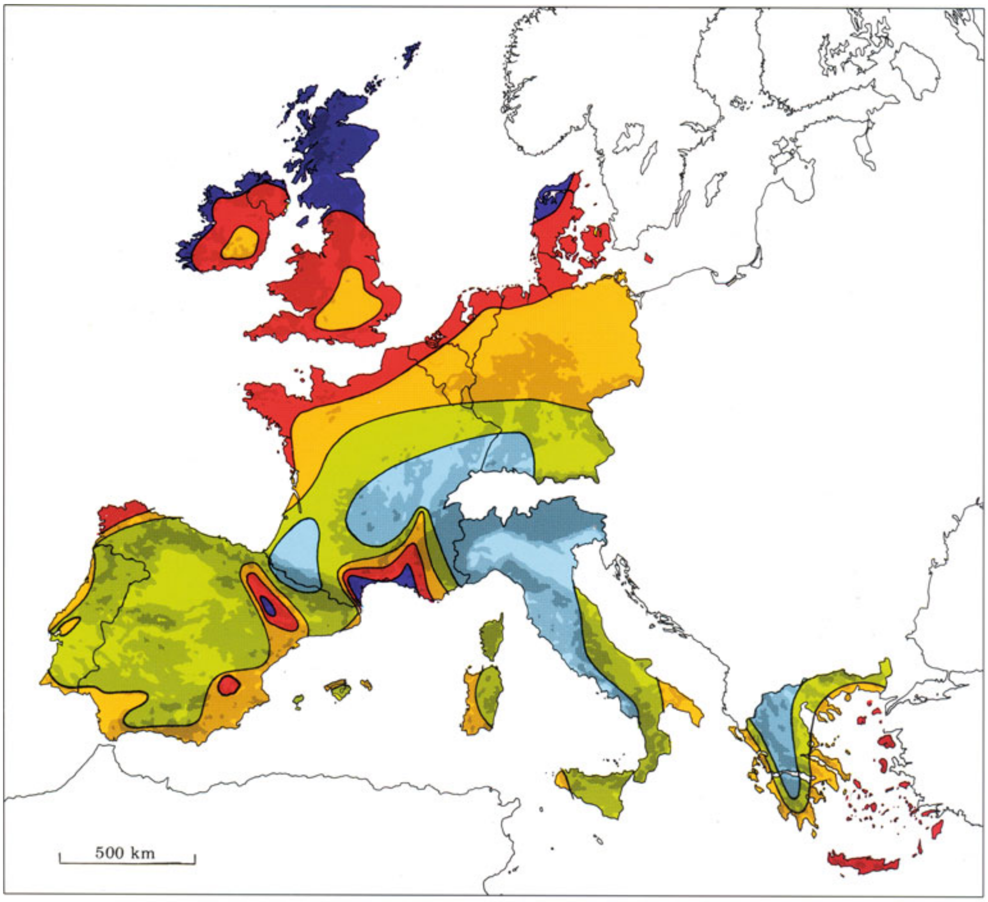
The European Wind Atlas is an example of the use of climatological data from the synoptic network to produce regional wind resource estimations. Another and even earlier example is the creation of the Danish Wind Atlas (1977–1980) [8], which actually builds on the application of a fundamental meteorological concept: The geostrophic wind and its climatology which was determined from long-term pressure measurements at about 55 synoptic stations in and around Denmark (43,000 km^2). The geostrophic wind climate was then used to estimate the wind distributions at a given height over a specific terrain, by means of the geostrophic drag law, which is a relation between the drag force on the surface, the roughness of the surface, and the geostrophic wind. The verification of the atlas was performed by estimating the wind climates of 12 specific sites in Denmark, where long-term wind measurements had been carried out. The procedure follows the right-hand side of Fig. 8, and the geostrophic wind climate is what in the figure is called the generalized regional wind climate. Hence, the development of "the wind atlas method" was initiated during the construction of the Danish Wind Atlas.

Both the Danish and the European Wind Atlas are examples on the use of climatological station of national synoptic networks, many of them located at airports and where routine observations were carried out by the meteorological and other public services. Many of these stations are part of the extensive Global Observing System. Data from satellites form the major part of the data from this system. Data from satellites are used intensively to map wind resources over the ocean. The principle is that the satellite can measure the speed of the capillary waves which then can be transformed into the wind speed [9].

The Global Data Archives

The data sources required for wind resource analysis can be broken into two parts, data about the atmospheric state and flow (known as reanalysis datasets), and data about Earth's surface topography. These two parts are shown below in two tables (Tables 1 and 2). For full information, use a search engine.

The Reanalysis Data Time series of the large-scale meteorological situation covering decades. These datasets have been created by assimilating measurement data from around the globe in a dynamical consistent fashion using large-scale numerical models. The primary purpose for the generation of the dataset is to provide a reference for the state of the atmosphere and to identify any features of climate change. For wind energy, the application of the dataset is as a long-term record of large-scale wind conditions.



Wind resources[1] at 50 metres above ground level for five different topographic conditions

	Sheltered terrain[2]		Open plain[3]		At a sea coast[4]		Open sea[5]		Hills and ridges[6]	
	m s^{-1}	Wm^{-2}	m s^{-1}	Wm^{-2}	m s^{-1}	Wm^{-2}	m s^{-1}	Wm^{-2}	m s^{-1}	Wm^{-2}
🟪	> 6.0	> 250	> 7.5	> 500	> 8.5	> 700	> 9.0	> 800	> 11.5	> 1800
🟥	5.0-6.0	150-250	6.5-7.5	300-500	7.0-8.5	400-700	8.0-9.0	600-800	10.0-11.5	1200-1800
🟧	4.5-5.0	100-150	5.5-6.5	200-300	6.0-7.0	250-400	7.0-8.0	400-600	8.5-10.0	700-1200
🟩	3.5-4.5	50-100	4.5-5.5	100-200	5.0-6.0	150-250	5.5-7.0	200-400	7.0- 8.5	400- 700
🟦	< 3.5	< 50	< 4.5	< 100	< 5.0	< 150	< 5.5	< 200	< 7.0	< 400

1. The resources refer to the power present in the wind. A wind turbine can utilize between 20 and 30% of the available resource. The resources are calculated for an air density of 1.23 kg m^{-3}, corresponding to standard sea level pressure and a temperature of 15°C. Air density decreases with height but up to 1000 m a.s.l. the resulting reduction of the power densities is less than 10%.
2. Urban districts, forest and farm land with many windbreaks (roughness class 3).
3. Open landscapes with few windbreaks (roughness class 1). In general, the most favourable inland sites on level land are found here.
4. The classes pertain to a straight coastline, a uniform wind rose and a land surface with few windbreaks (roughness class 1). Resources will be higher, and closer to open sea values, if winds from the sea occur more frequently, i.e. the wind rose is not uniform and/or the land protrudes into the sea. Conversely, resources will generally be smaller, and closer to land values, if winds from land occur more frequently.
5. More than 10 km offshore (roughness class 0).
6. The classes correspond to 50% overspeeding and were calculated for a site on the summit of a single axisymmetric hill with a height of 400 metres and a base diameter of 4 km. The overspeeding depends on the height, length and specific setting of the hill.

Meterology and Wind Power. Figure 10
Over view map from the European Wind Atlas – the large-scale variation of the European wind climate. By means of the legend, a range for mean wind speed and mean wind energy at the height of 50 m can be estimated for five topographical conditions. This map sends the important message to politicians, decision makers, and the general public, that even in low wind areas it is possible to find locations for wind farms with a good wind resource

Meterology and Wind Power. Table 1 Global reanalysis datasets (Courtesy Jake Badger)

Product	Model system	Horizontal resolution	Period covered	Temporal resolution
ERA Interim reanalysis	T255, 60 vertical levels, 4DVar	~0.7° × 0.7°	1989–present	Three-hourly
NASA – GAO/ MERRA	GEOS5 data assimilation system (Incremental analysis updates), 72 levels	0.5° × 0.67°	1979–present	Hourly
NCAR CFDDA	MM5 (regional model) + FDDA	~40 km	1985–2005	Three-hourly
CFSR	NCEP GFS (global forecast system)	~38 km	1979–2009 (& updating)	Hourly

Meterology and Wind Power. Table 2 Topographical datasets (Courtesy Jake Badger)

Product	Product description
Elevation – Shuttle Radar Topography Mission (SRTM), version 2.1, released 2009	SRTM data are available as grid point spot heights with a resolution of 1 arc-second (continental USA) and 3 arc-seconds (from 56°S to 60°N, 80% of the Earth's land surface). Derivative datasets exist, where data voids have been filled. The SRTM datasets are used extensively for wind resource assessment already and they are easy to download, process, and transform (e.g., to height contour maps)
Elevation – ASTER Global Digital Elevation Model (ASTER GDEM), version 1, released 2009	ASTER data are available as grid point spot heights with a resolution of 1 arc-second from 83°S to 83°N (99% of the Earth's land surface). Less experience exists using ASTER data for wind resource assessment but they are in principle easy to download, process and transform
Coastline contours – SRTM Water Body Data (SWBD)	Version 2 of SRTM also contains the vector coastline mask derived by NGA during the editing, called the SRTM Water Body Data (SWBD), in ESRI Shapefile format. SWBD data covers the Earth's surface between 56°S and 60°N; the rest of the world is available at lower resolution through the Coastline Extractor hosted by NOAA
Land cover – ESA GlobCover, version 2.1, released 2008	GlobCover is the highest resolution (300 m) global land cover product ever produced and it is made available to the public by the European Space Agency (ESA). The GlobCover land cover map is compatible with the UN Land Cover Classification System (LCCS)
Land cover – regional databases	Several regional and national land cover datasets exist which may be more detailed and sometimes more readily applicable to microscale flow modeling. As an example, the European Environmental Agency (EEA) has produced vector and raster land cover databases – CORINE – for EC Member States and other European countries

Topographical Datasets Modeling of the wind resource requires detailed information on the terrain elevation, water body distribution, and land cover of the terrain. Several global or near-global datasets are available that are used for this purpose.

Assessing Local Wind Conditions for Design and Turbine Safety

The previous paragraphs have focused on the part of the wind conditions which specifically are directed toward the determination of the wind resource – site dependent or regionally dispersed. Clearly, here the aim is to find the most favorable windy sites or regions for the employment of wind energy. But then the question remains: Are the wind conditions also favorable for operating wind turbines with respect to structural loading and damage to structure and components?

Wind turbine manufacturers offer a range of turbines designed for various wind regimes, characterized in terms of turbulence, mean wind, and extreme wind. The challenge is to select the right turbine for the right site and in order to do this, wind farm developers need to perform verification of design conditions and turbine safety. In this paragraph, a brief introduction to the verification of site conditions is given in accordance with the IEC 61400–1 standard.

Design Standards and Turbine Classes

The International Electrotechnical Commission (IEC) provides standards for electrical equipment including wind turbines. The main principle for ensuring structural safety of wind turbines in the IEC 61400–1 standard, is to design a wind turbine type to a set of design conditions defined for a specific turbine class, and later to verify that site conditions match the turbine class. Turbine classification and design are the responsibility of manufacturers, and site assessment is the responsibility of project developers [10].

To meet the class requirements according to IEC 61400–1, a turbine must be demonstrated to be safe for a range of predefined load cases. In this context, wind conditions are specified by generic models scaled by a reference wind speed V_{ref}, reference turbulence intensity (TI), and turbine hub height. Wind speed and TI are defined for hub height and 10-min averaging periods. The reference wind is the extreme wind with 50-year recurrence, and the reference TI is the average TI at a mean wind speed of 15 m/s. These key wind characteristics at hub height define the wind turbine classes in IEC 61400–1. The wind turbine classes are intended to cover most applications, but the particular external conditions defined for classes I, II, and III are neither intended to cover offshore conditions nor wind conditions experienced in tropical storms such as hurricanes, cyclones, and typhoons (Table 3).

Meterology and Wind Power. Table 3 Basic parameters for wind turbine classes in IEC 61400-1

Wind turbine class		I	II	III	S
V_{ref}	(m/s)	50	42.5	37.5	Values specified by the designer
A	I_{ref} (–)	0.16			
B	I_{ref} (–)	0.14			
C	I_{ref} (–)	0.12			

The turbine class is designated by a roman number (I, II, or III) which refers to the reference wind speed and a suffix letter (A, B, or C) which refers to wind and turbulence, respectively. Thus, a turbine for high wind and medium turbulence is categorized as class I_B.

Atmospheric TI has a variation which is mainly caused by variable atmospheric stability and topographic variations for various wind directions. In general, mean and scatter of the TI decrease with wind speed, though offshore the mean TI will increase when the wind begins to build up waves.

Material damage is not linearly proportional to turbulence, so the IEC standard specifies a design curve, which corresponds to the 90th percentile of the distribution for both extreme and fatigue load estimation. Dynamic loads in operation are estimated by a chain of calculations, which first simulate time series of wind with specified turbulence statistics; then simulate series of aeroelastic responses; and finally evaluate fatigue damage and load extremes.

The simulations are repeated for a range of wind speeds, and the overall fatigue damage is found by an integral weighted by wind speed probability. Vertical wind shear also contributes to fatigue loads, and a somewhat conservative value is prescribed for the aeroelastic simulations. The turbine is considered safe when neither the accumulated material damage over 20 years nor the extreme loads lead to failure. Safe response to severe gust, sudden wind shear, or sudden change of wind direction is verified by special load cases.

As stated above, turbine classification and design are the responsibility of manufacturers, and site assessment is the responsibility of project developers. The purpose of a site assessment is to verify that actual site conditions or resulting loads are less severe than assumed in the design. In any case, the actual site conditions must be estimated.

There are a number of criteria which apply for the site conditions:

- The 50-year extreme wind must not exceed V_{ref}.
- Flow inclination must not exceed ±8°.
- Average wind shear must neither be negative nor exceed the design wind shear.
- The wind speed distribution must not exceed the distribution of the turbine class between 0.2 and 0.4 V_{ref}.
- Effective TI must not exceed the TI of the turbine class between 0.2 and –0.4 V_{ref}.

Effective TI is a uniform condition with the same material damage as variable turbulence in winds from various directions. It facilitates comparison of variable field conditions with the uniform TI assumed for the turbine design process. The site-assessment criteria apply to individual turbine sites.

The verification is commonly done by a comparison of the design conditions with conditions at the actual site, determined by a combination of measurements, models, and rules for increasing design parameters, for example TI, if this parameter is not measured. Alternatively, the loads originating from the actual site conditions must be demonstrated to be below the design loads.

An important design parameter is the reference wind speed V_{ref} which is compared with the extreme wind estimated for the site. Extreme winds occur as the highest wind speed during individual storms and they are considered mutually independent. It is possible to fit an extreme wind model to observed extreme winds, and extrapolate to a 50-year recurrence period. Project developers often prefer data from local met-masts installed for resource estimates, but reasonably accurate extrapolation to 50 years obviously requires quite a long time series. The IEC standard recommends a minimum of 7 years, which exceeds most wind energy measurement campaigns. Therefore, recourse is often taken to use data from a nearby reference station. Winds will differ between the reference site and the turbine sites, so it is necessary to correct the observations by a microscale flow model before making the statistical analysis. This can be done using the wind atlas method as previously described.

The measurements at the reference station are then converted to winds over flat terrain with uniform standard surface roughness, and the result is a regional extreme wind atlas. Winds at the potential wind farm sites are then predicted in a second model domain by similar but reverse corrections on the atlas data as it is depicted in Fig. 8.

The reference station could be part of the global meteorological observation network. As a substitute, in case no suitable data are available and with some reservations, a global reanalysis set can be processed for prediction of regional extreme wind climates by means of a mesoscale–microscale combination.

Future Directions

To put "payed" to the problem of determining the right wind conditions for calculating wind resources and structural loads, three things are necessary:

- A body of well-measured data
- A set of models that can reproduce the data
- A theory based on first principles that fully connects measurements and models

Therefore, any advances in this field require an effort on all three components and the fast development in measuring and information technology will continue to contribute. But only advances in theoretical and numerical description of the dynamics of atmospheric motions and the understanding of the interaction with the structural dynamics of wind turbines can secure a significant advancement.

Bibliography

Primary Literature

1. IEC (2005) Wind turbines. Part 1: design requirements, International Standard 61400–1. International Electrotechnical Commission, Geneva
2. Troen I, Petersen EL (1989) European wind atlas. Risø National Laboratory. Published for the EEC, Roskilde, 688 pp (Also in Italian, French, Spanish and German)
3. Krohn S, Poul-Erik Morthorst P-E, Awerbuch S (2009) The economics of wind energy. European Wind Energy Association, Brussels, 155 pp
4. Petersen EL, Mortensen NG, Landberg L, Højstrup J, Frank HP (1998) Wind power meteorology. Part 1: climate and turbulence. Wind Energy 1:2–22

5. Petersen EL, Mortensen NG, Landberg L, Højstrup J, Frank HP (1998) Wind power meteorology. Part 2: siting and models. Wind Energy 1:55–72
6. Kaimal JC, Finnigan JJ (1994) Atmospheric boundary layer flows. Their structure and measurement. Oxford University Press, New York, 289 pp
7. Warner TT (2011) Numerical weather and climate prediction. Cambridge University Press, Cambridge, UK, 526 pp
8. Petersen EL, Troen I, Frandsen S, Hedegaard K (1981) Danish wind atlas. A rational method of wind energy siting. Risø-R-429. Risø National Laboratory, Roskilde, 229 pp
9. Badger M, Badger J, Nielsen M, Hasager CB, Pena Diaz A (2010) Wind class sampling of satellite SAR imagery for offshore wind resource mapping. J Appl Meteor Climatol 49(12):2474–2491, American Meteorological Society
10. Nielsen M, Jørgensen HE, Frandsen ST (2009) Wind and wake models for IEC 61400–1 site assessment, 2009. In: European wind energy conference, Marseilles. 6 p

Books and Reviews

EWEA (2009) Wind energy, the facts. European Wind Energy Association, Brussels

In general: The conference proceedings published yearly by the European Wind Energy Association (EWEA), The American Wind Energy Association (AWEA) and the British Wind Energy Association (BWEA)

Jensen NO, Petersen EL, Troen I (1984) Extrapolation of mean wind statistics with special regard to wind energy applications. Report No. WCP–86. World Climate Programme, WMO, Geneva, 85 pp

Wyngaard JC (2010) Turbulence in the atmosphere. Cambridge University Press, Cambridge, UK, 393 pp

Microbial Risk Assessment of Pathogens in Water

Gertjan Medema
Water Quality & Health, KWR Watercycle Research Institute, Nieuwegein, The Netherlands
Sanitary Engineering, Delft University of Technology, Delft, The Netherlands

Article Outline

Glossary

Dose-response assessment The determination of the relationship between the magnitude of exposure (dose) to a microbiological agent and the severity and/or frequency of the associated adverse health effects (response).

Exposure assessment Qualitative and/or quantitative evaluation of the likely intake of microbial hazard via all relevant sources or a specific source.

Exposure Concentration or amount of an infectious microorganism that reaches the target population, or organism usually expressed in numerical terms of substance, concentration, duration, and frequency.

HACCP: Hazard Analysis Critical Control Point A system that identifies, evaluates, and controls hazards that are significant for water safety.

Hazard A biological agent with the potential to cause an adverse health effect.

Hazard identification The identification of microbiological and biological agents capable of causing adverse health effects that may be present in water.

Hazardous event An event that may lead to the presence of a hazard in drinking water.

Health effects Changes in morphology, physiology growth, development or life span of an organism, which results in impairment of functional capacity or impairment of capacity to compensate for additional stress or increase in susceptibility to the harmful effects or other environmental influences.

Infection Colonization of a human (tissue) by a microorganism.

Infectious disease Colonization by a pathogenic microorganism leading to overt symptoms of disease.

Pathogen A microorganism capable of causing disease.

QMRA Quantitative Microbial Risk Assessment.

Risk assessment A scientifically based process consisting of the following steps: (1) hazard identification, (2) exposure assessment, (3) effect assessment, and (4) risk characterization.

Risk characterization The qualitative and quantitative estimation, including attendant uncertainties of the probability of occurrence and severity of known or potential adverse health effects in a given population based on hazard identification, hazard characterization, and exposure assessment.

Risk The likelihood of occurrence of an adverse health effect consequent to a hazard in drinking water.

Uncertainty Lack of knowledge about specific factors, parameters, or models. Uncertainty includes parameter uncertainty (measurement errors, sampling errors, systematic errors), model uncertainty (uncertainty due to necessary simplification of real-world processes, mis-specification of the model structure, model misuse, use of inappropriate surrogate variables), and scenario uncertainty (descriptive errors, aggregation errors, errors in professional judgment, incomplete analysis).

Variability Intrinsic heterogeneity in a population, process, or parameter.

Water Safety Plan (WSP) A management plan developed to address all aspects of water supply that are under the direct control of the water supplier focused on the control of water production, treatment, and distribution to deliver drinking water.

Definition of the Subject

Water can transmit infectious diseases. Water can be transport vehicle. A range of pathogenic microorganisms is shed into the water cycle by infected hosts (man or animal) and transported to new hosts by the water cycle. Water can also be a niche for (opportunistic) pathogens. These pathogens grow in water ecosystems (natural or man-made) and may infect humans that come into contact with this water. Management of the risk of waterborne disease transmission requires knowledge about the nature of the pathogens, their potential growth, fate and transport in the water cycle, the routes of exposure to humans and the health effects that may result from this exposure in the human population, as well as the effect of potential mitigation measures. The challenge is to combine all this knowledge into information that risk managers can use. Quantitative Microbial Risk Assessment (QMRA) has developed as a new scientific discipline over the last 2 decades as a transparent, science-based approach that allows the risk manager to use the best available scientific evidence as basis for risk management decisions.

Introduction

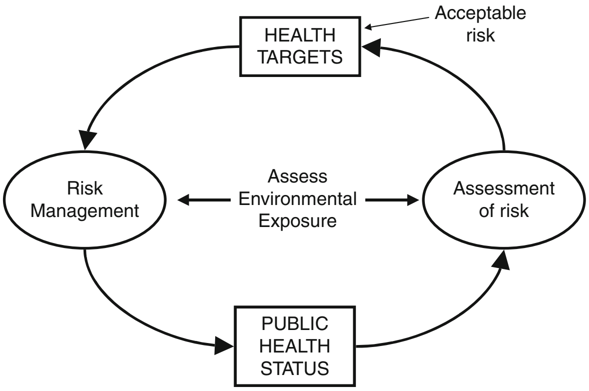
We run risks. We always have. From being eaten by lions, being slaughtered by a rivaling tribe, to being hit by a car. A principal objective of decision making has always been to reduce risks. From avoiding lions, building walls around cities to regulating traffic. To make wise decisions, it is important to have good information about risks. Risk assessment aims to aid decision makers by collating and evaluating this type of information. Risk assessment is increasingly applied in our society, for a wide range of activities: economy, finance, insurances, traffic, infrastructure, health, and environment. What all these activities have in common is that we want to reduce our risk and need to spend resources on mitigation measures. Our resources are limited, so we need to allocate them wisely and proportionally. Risk assessment helps to keep proportions. Risk assessment as a formal discipline has emerged after World War II, paralleling the developments in air and road traffic, the nuclear power and chemical industries and the need to improve the safety of these activities. The process of risk assessment tries to determine the probability that a hazardous event will occur and the probable magnitude of the adverse effects that such an event will have. In the Netherlands, where a substantial part of the country lies below sea level and is protected against flooding by dikes, the height and strength of the dikes are based on assessing the probability of a storm event and the probable magnitude of the adverse effects of flooding part of the country.

In the health and environment arena, risk assessment science has developed over the last few decades. In environmental health, scientists try to establish the probability of exposure of humans to toxic chemicals or pathogens and the probable magnitude of the health effects of this exposure. Risk assessment has become a dominant tool in environmental policy-making, For chemical risks, this is well established (although not without debate [52]). Regulatory agencies are using chemical risk assessment to set standards for toxic chemicals in water. For risks of pathogenic microbes via water, the use of risk assessment was first proposed in the early 1990s [60]. The World Health Organization has been instrumental in the introduction of microbial risk assessment as a basis for safety management of the water we use for drinking, recreation, and food crop irrigation [73, 74].

The Safe Water Framework

An international group of experts, assembled by the World Health Organization, discussed the approach to assess and manage the health risk of pathogenic microorganisms in drinking water, recreational water, and wastewater reuse [7]. This group agreed that future guidelines for safe water and sanitation should integrate risk assessment and risk management into a single framework, the Safe Water Framework. The simplest form of the framework is shown in Fig. 1.

The risk that is assessed and managed in this approach is a health risk. It is clearly an iterative cycle in which risk assessment is a basis for decision making in risk management. The four steps of the cycle are described in the next paragraphs, using drinking water safety as an example. In the World Health Organization (WHO) guidelines for the safe use of wastewater, excreta, and grey water [74], these same steps are used for assessing and managing the risk of these water systems.



Microbial Risk Assessment of Pathogens in Water. Figure 1
Safe Water Framework for integrated risk assessment and risk management

Health Targets

Health targets are benchmarks for water suppliers, set by the regulator as part of their health policy. Health targets for drinking water are traditionally strict because of the large impact of contaminated tap water and the basic need for safe drinking water. That leads to the question of what level of health risk through drinking water could be tolerated, given the overall health status of the consumer population and the contribution of drinking water to the overall health risk of this population in relation to other routes of exposure, such as food, person-to-person or animal contact, recreational water, etc. This is a question that typically needs answering on the level of the regulator, who can translate this information into a health target for drinking water, considering other factors such as relative contribution of drinking water–transmitted disease to the overall health burden and the economic climate.

The health target is the level of tolerable risk for drinking water, which could be expressed as the tolerable risk of infection through drinking water (i.e., risk of infection $<10^{-4}$ per person per year [61]) or the tolerable amount of disease burden (i.e., $<10^{-6}$ disability adjusted life years per person per year [31, 73]). The health target could be translated into water quality targets for pathogens (analogous to the toxic chemicals). In the latter case, rather than producing a standard and monitoring requirement for all pathogens that could be transmitted through drinking water, the use of a suite of "index pathogens" is advisable. Establishment of adequate control against this suite of pathogens should offer protection against the other known (and even unknown) pathogens.

It is emphasized that the health targets may be different in different health status situations. The question of what is a tolerable level of risk is a judgment in which the society as a whole has a role to play; the decision on the cost-benefit is for each country to decide [71, 73]. It is important that health-based targets, defined by the relevant health authority, are realistic under local operating conditions and are set to protect and improve public health. Health-based targets underpin development of Water Safety Plans [73] and provide information with which to evaluate the adequacy of existing installations, and assist in identifying the level and type of inspection and analytical verifications appropriate.

Risk Management

Managing the safety of drinking water has been the core business of water supply companies for more than a century. Over this period, risk management has evolved into a culture, with codes and specifications of good practice. In the last few decades, quality management systems have been used in the water industry to formalize these practices. Currently, water suppliers in several European Union (EU) countries are using a Hazard Analysis and Critical Control Points (HACCP) based approach for management of (microbiological and other) risks. The basic principles of HACCP are to understand the system and the hazards/hazardous events that may challenge the system and their (health) priority and to ensure that control measures are in place and functioning. HACCP-based systems typically focus on good practice and even more specifically on ensuring that good practice is maintained at all times. HACCP fits within existing quality management systems (i.e., ISO 9001 c.s.). HACCP is the risk management tool that is used in food safety. The Codex Alimentarius (FAO/WHO code for food safety) defines HACCP as a system that identifies, evaluates, and controls hazards that are significant for food safety [10]. The HACCP system is well established in the food industry.

Although there are many aspects of drinking water that are similar to food, there are also differences. Based on experiences of water suppliers with HACCP, the HACCP system has been refined and tailored for application in drinking water abstraction, treatment, and distribution in WHO's Water Safety Plan. The Water Safety Plan is described in the third revision of the Guidelines for Drinking Water Quality [73].

The principal components of the Water Safety Plan are:

System assessment to determine whether the water supply chain (from source through treatment to the point of consumption) as a whole can deliver water of a quality that meets the above targets.

Operational monitoring of the control measures in the supply chain that are of particular importance in securing drinking water safety.

Management plans documenting the system assessment and monitoring, and describing actions to be taken in normal operation and incident conditions, including upgrade and improvement documentation and communication.

In the Water Safety Plan, the risk assessment question: "Do we meet the health target?" is answered in the *System Assessment* and the risk management questions "How do we ensure and demonstrate that we always meet the target?" and "How do we respond to incidents?" are answered in the *Operational monitoring of control measures* and the *Management plans*.

For an overview of the Water Safety Plan and its context, the reader is referred to the WHO GDWQ and the Water Safety Plan guidance documents that are published on the website of WHO Water, Sanitation, and Health.

Public Health Status

The primary objective of drinking water safety management is the adequate protection of public health. The incidence of waterborne illness in the population or the occurrences of waterborne outbreaks are direct triggers for curative risk management. A more preventative incentive for assessing the water-related health risks and the installation of risk management is to demonstrate that the water supply is providing an adequate level of protection of public health.

The installation of health targets in national legislation and the risk management actions of water utilities should result in an improvement of the status of public health. Without addressing this, it is impossible to see if the health targets set and risk management actions taken are effective and if money spent for improving water supply results in a relevant health gain. This step in the process is the place where the health risk of drinking water can be compared to other routes of exposure and to other health risks. It allows comparison of the effort and resources put into the provision of safe drinking water versus resources allocated to manage other health risks.

The risk assessment and management framework is a circular process that can be run in an iterative manner. This fits well with the incremental nature of health decision-making, the efficient use of scarce resources, and the increase of information each time the circle is completed.

Risk Assessment

Risk assessment is used to answer the question: “Is my system able to produce and deliver drinking water that meets the health targets?” The risk assessment process requires quantitative information about the exposure of drinking water consumers to pathogens. This is provided by exposure assessment, one of the components of risk assessment. Quantitative information about pathogens in water sources, their removal by treatment and protection of the distribution network and drinking water consumption is collected and translated into an estimate of the exposure of consumers to pathogens through drinking water. To complete the risk assessment, the potential effect (the risk) of pathogen exposure is estimated through known dose-response models. As will be indicated later, the exposure assessment also provides valuable information to aid risk management in the prioritization of control measures.

An important question in risk management, especially in settings with an already high standard of drinking water safety, is “How far do we need to go with control measures?” This is an optimization that weighs the safety of the consumer against the costs of drinking water.

Quantitative microbial risk assessment (QMRA) can provide an objective and scientific basis for risk management decisions. Water utilities can use QMRA to assess whether they meet the health targets with their water treatment, storage, and distribution systems. This also provides the information to set the critical limits in the Water Safety Plans to ensure good performance. Good performance can now be based on a quantitative assessment of the contribution of the Critical Point (such as a disinfection or filtration process) to the overall safety, and limits can be set to ensure that the multiple barrier chain of water collection, treatment, and distribution as a whole does meet the target.

Risk assessment and risk management should not be regarded as two separate steps in the harmonized framework. To answer the question “Which control measures should be put in place to meet the target?” both the HACCP-based system and quantitative risk assessment provide valuable input: the hazardous events, the most important barriers in the system, the contribution of each of the barriers, target levels for control, the occurrence of weak elements in the chain, the quality of the available information, etc.

Quantitative Microbial Risk Assessment

Quantitative Microbial Risk Assessment (QMRA) is derived from the chemical risk assessment paradigm that encompasses four basic elements:

- A characterization of the problem, including the hazard
- Exposure assessment
- Effect assessment (dose-response)
- Risk characterization

Several QMRA frameworks have been published, such as the generic International Life Sciences Institute (ILSI) framework [8]. Here, most attention is given to exposure assessment and risk characterization of pathogens in drinking water. Therefore, the generic ILSI QMRA framework is expanded to highlight the elements that are important for exposure assessment and risk characterization in drinking water, and put in the overall WHO Safe Water Framework (Fig. 2).

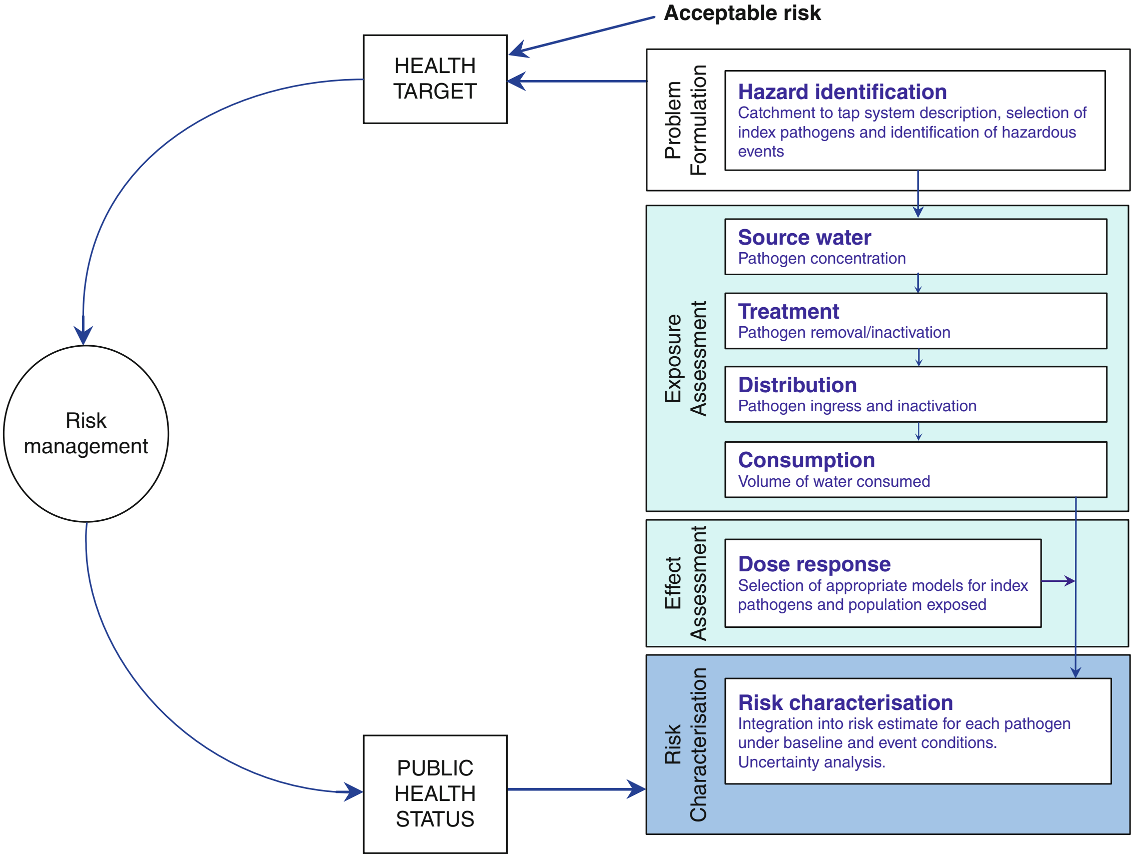
Element 1. Problem Formulation and Hazard Identification

This is the initializing phase of QMRA to establish which specific questions need to be addressed. The scope and the boundaries of the QMRA process are determined in this phase. This requires communication between the risk managers (regulators, public health agencies, water utilities) and the risk assessors. The basic question to QMRA is: “Is my system able to meet the health targets?”

To conduct a QMRA, a good description of the system under evaluation is necessary and the hazards and hazardous events need to be identified.

Step 1. Description of the System from Source to Tap The system for water treatment from catchment to tap is described, identifying the principal control elements and strategies.

Step 2. Hazard Identification Hazard identification is the identification of the microorganisms within the system boundaries that cause human illness, the processes by which each microorganism causes illness and


Acceptable risk
HEALTH TARGET
Problem Formulation
Hazard identification
Catchment to tap system description, selection of index pathogens and identification of hazardous events
Exposure Assessment
Source water
Pathogen concentration
Treatment
Pathogen removal/inactivation
Distribution
Pathogen ingress and inactivation
Consumption
Volume of water consumed
Risk management
Effect Assessment
Dose response
Selection of appropriate models for index pathogens and population exposed
Risk Characterisation
Risk characterisation
Integration into risk estimate for each pathogen under baseline and event conditions.
Uncertainty analysis.
PUBLIC HEALTH STATUS

Microbial Risk Assessment of Pathogens in Water. Figure 2
The steps of quantitative microbial risk assessment in the Safe Water Framework

the type of illness(es) caused, and the identification of possible transmission routes and the significance of these routes [26]. QMRA is usually focused on a specific transmission route, in this example drinking water from a surface water source.

The ideal QMRA does not focus on a single pathogen only, but on a suite of "index pathogens" that cover the range of health risks and control challenges for the particular water supply system defined. Adequate control of these index pathogens implies that the health risk of other known pathogens is also adequately controlled by the system and that the system also offers protection against unknown pathogens.

Hazard identification consists of the following steps:

Description of the characteristics of the pathogens, especially those related to waterborne transmission (survival in water, resistance to treatment, etc.).

Description of what is known about the transmission routes of these pathogens and specifically what is known about waterborne transmission, the causes of waterborne outbreaks, and the relative significance of waterborne transmission compared to other routes.

Description of the illness (type, duration, incubation time, etc.) caused by the pathogens in the risk assessment, and available information about sequellae.

Description of what is known about protective immunity and secondary transmission.

Step 3. Description of Hazardous Events In many cases, the majority of the risk is not determined during the normal (baseline) situation, but during hazardous events, such as rainfall leading to a high load of pathogens in source waters, or treatment failure or distribution network failure (or combinations thereof). It is therefore important to ensure that these hazardous events are incorporated in the QMRA, or that a separate QMRA is conducted to determine the (health) significance of the event.

Element 2. Exposure Assessment

Exposure assessment is the quantitative assessment of the probability that drinking water consumers ingest pathogens. A QMRA of drinking water usually requires the assessment of the levels of pathogens in source water and the changes to these levels by treatment, storage, and distribution, and finally the volume of water consumed.

Step 4. Assess Pathogen Occurrence in Source Water Collect information about the occurrence of pathogens in source water. This is preferably based on a catchment survey, identifying the principal sources of contamination of the catchment and the conditions that may lead to peak events in source water, such as heavy rainfall or resuspension of sediments. Pathogen monitoring in source water can be carried out, using the information of the catchment survey, which needs to include assessment of peak events. The pathogen detection methods are ideally targeted to viable and infectious pathogens. The performance characteristics of the available detection methods for pathogens can have implications for the applicability of the data in risk assessment. These should be identified and evaluated in (the early stages of) the risk assessment process.

Step 5. Assess the Elimination of Pathogens During Treatment Collect information about the removal or inactivation of pathogens during drinking water treatment processes. Ideally, data on removal of pathogens at full scale are used. In practice, however, several other sources of data have to be used to estimate pathogen removal, such as pathogen data of pilot or lab scale systems or data on model parameters (indicator bacteria, phages, spores, particles, etc.) on full, pilot, or lab scale.

The efficacy of treatment processes may vary, depending on feed water composition, operational control, temperature, etc. Moments or periods of poor or suboptimal performance are hazardous events and hence most significant for risk assessment.

Step 6. Assess the Changes in Water Quality During Storage and Distribution Determine the likelihood of recontamination of stored and distributed water (e.g., by the *E. coli* monitoring of water in these reservoirs and pipes or loss of disinfectant residual) and the significance of these contamination events. In well-maintained piped supplies, recontamination events are rare and could be regarded as a result of a hazardous event (heavy rainfall, cross-connection, poor hygiene during repairs, etc.). In other piped and non-piped settings, recontamination events are common and may dominate the health risk.

Step 7. Consumption of Drinking Water The other component of exposure assessment is the volume of water consumed by the population. Not only the average volume of water consumed is important, also the person-to-person variation in consumption behavior and especially consumption behavior of risk groups (in terms of sensitivity to infection or high level of consumption) is relevant. The available data suggest there is considerable difference between drinking water consumption within the population. This variation needs to be captured and incorporated in the risk assessment. Household treatment/point-of-use devices affect the exposure. Hence, consumption data should be on consumption of drinking water without further treatment, such as heating or filters and include water that is drunk directly, but also cold tap water used for food preparation, ice, etc.

Step 8. Dose (Exposure) Estimation Dose (or exposure) is the number of pathogens consumed per unit time. The information obtained in all steps of the exposure assessment needs to be combined into an estimate of the ingested dose. This is preferably a stochastic estimation, including the variability and uncertainty in all steps of the exposure assessment.

Element 3. Effect Assessment

The effect assessment is the determination of the health outcomes associated with the (level of) exposure to waterborne pathogens.

Step 9. Dose-Response Data Dose-response characterizes the relation between dose magnitude, infectivity, and quantitative health effects to an exposed population. The microbial dose-response analysis records the incidence of a particular effect against dose of the agent. In most cases, this particular effect is infection, rather than symptoms of illness. For *Cryptosporidium parvum* for instance, there is a clear relation between ingested dose and the probability of infection, but not between dose and symptoms of intestinal illness.

Although the dataset is increasing, the number of dose-response studies with human volunteers is limited. Of most pathogens, only one or a few strains are tested in healthy adult volunteers. Information about strain-to-strain variability and the influence of the immune response of the hosts is still limited.

There are several dose-response models available and the type of model can have a very significant impact on the response that is attributed to exposure to low doses. The models and their limitations should be well understood when applying these in QMRA. Synergistic effects between pathogens are not incorporated in the current models.

Step 10. Host Characterization For infectious diseases, the host susceptibility plays an important role in the health outcome of exposure. Exposure of persons with protective immunity will result in lower health outcomes than exposure of risk groups. During "Host Characterization" the characteristics of the potentially exposed populations that are suspected for susceptibility to a particular pathogen are evaluated.

Step 11. Health Outcome Until now quantitative microbial risk assessment has been primarily focused on estimating the risk of infection. The relation between ingested dose and infection is relatively well defined, while the relation between dose and other health outcomes (illness, sequellae) is not available or less clear. This is one of the reasons why it is difficult to establish a direct relation between QMRA (on probability of infection) and epidemiological data (on symptoms of disease). The use of the risk (or probability) of infection is justified by the degree of conservatism in using infection as an endpoint and the inability to quantify the risk of more susceptible subpopulations [43].

However, waterborne diseases differ in nature, severity, and duration. A metric that takes into account the overall health burden of waterborne diseases is necessary. Ideally, this metric can also be used to describe the burden of the disease of chemical compounds, such as carcinogens, so all health risks can be weighed on the same scale.

In the new WHO guidelines for Drinking-Water Quality (GDWQ), the concept of Disability Adjusted Life Years (DALY) [31] is introduced as burden of disease metric in the drinking water guidelines.

The basic principle of the DALY approach is to weigh each health effect for its severity with (usually) death as the most severe outcome, multiply this weight with the duration of the health effect ("duration" of death being the remaining group life expectancy), and with the number of people in a population affected by the particular outcome. Summarizing all the health outcomes caused by a certain agent results in an estimate of the burden of disease attributable to this agent.

To be able to use DALYs in the QMRA, ideally the relation between exposure (dose) and different health outcomes is known. In the absence of sufficient data (which is usually the case), the dose-response relation for infection (as the first step of the disease process) can be combined with data on the fraction of the exposed population falling ill from exposure (for instance, from attack rates in waterborne outbreaks) and data on the fraction of the ill population that contract more severe health outcomes (from health surveillance data).

Element 4. Risk Characterization

In the process of risk characterization, the information obtained in the exposure assessment and the effect assessment are integrated to obtain a risk estimate. This can be done as a point estimation: a point estimate of exposure can be entered into the dose-response relation to compute a point estimate of the risk of infection. The point estimate can be the "best" estimate, to obtain a measure of central tendency of the risk. In the case of computing various risk scenarios, the computed point estimates give a quantitative estimate of the consequences of the circumstances that produce a risk scenario.

An approach that allows the incorporation of the variability and uncertainty in the steps of the risk assessment chain is promoted by [23, 66]. This encompasses the characterization of the distribution of all data used for risk assessment and to combine these distributions into a distribution of the computed risk, for instance, by Monte Carlo analysis. This approach not only provides the risk manager with important information about the (un)certainty of the risk estimate, but also with the relative contribution of the uncertainty and variability in all steps of the risk assessment. It therefore guides the risk manager to the most appropriate options for efficiently minimizing the risk and the most significant research items to reduce the overall uncertainty of the risk estimate.

With high-level water supply, the baseline risk is usually very low. Under such conditions, hazardous events, such as peak contamination in the source water, treatment failure and especially the combination thereof and contamination events in the distribution network, are responsible for the majority of the risk. Most waterborne outbreaks have been traced to a combination of hazardous events [35] and it is likely that many events result in the presence of pathogens in tap water and hence the transmission of disease. Wherever possible, identify and evaluate these events separately in QMRA to understand the significance of these events. Analysis of events also brings forward opportunities for optimization of the system to prevent these events from occurring or reduce their impact on health.

Tiered Approach

Risk assessment is well suited for a tiered approach and this is also commonly used in risk assessment practice, both in human health risk assessment and in ecological risk assessment. The tiered approach allows an effective interaction between risk assessment and risk management, starting with a crude risk assessment, usually based on limited information to determine the urgency of the perceived problem, to prioritize the risk of different water supply sites or scenarios, and to determine the need of a more detailed study for a particular situation. This allows the effective allocation of resources to the sites or situations that give rise to the highest risk. There is no strict definition of the tiers, only that the initial QMRA is usually generic and simple and the specificity and complexity increase in subsequent tiers.

The most basic (but also most important) QMRA is a screening-level study. Starting with whatever information is available, a crude first evaluation is made. Usually, the available information is not specific to the system that is studied, but has to be extrapolated from the available scientific literature. So, in its simplest form, a QMRA can be performed with only a generic description of the water supply system.

The screening-level assessment may show that the risks are negligible, without much scientific doubt. In that case, the screening-level risk assessment can be used to demonstrate the safety of the system. Setting up a more detailed study is not warranted. Or the screening-level risk assessment may highlight that the risk is unacceptably high, again without much scientific doubt. Such a screening-level risk assessment is also very useful in comparing different scenarios for risk management, for example, different water treatment options.

If the outcome of the screening-level risk assessment is that there may be a health risk that is not negligible, there is an incentive for a next iteration of the risk assessment, the collection of site-specific data, for instance, on the presence of *Cryptosporidium* in the source water or catchment. The QMRA is repeated with the new, site-specific information. The options for the outcome of this second-level QMRA are the same as for the first iteration. In general, a result of any risk assessment is the identification of which information is missing and the prioritization of research needs [21].

The screening-level risk assessments usually work with point estimates of risk. The tendency is to use conservative or worst-case estimates, to “be on the safe side.” But worst-case estimates, by nature, may overestimate the risk and it is not clear to the risk manager what the uncertainty of the calculated risk is, only that the uncertainty will be toward the lower risk values (the nature of a worst-case assumption). More helpful for the risk manager is to provide a range of risks (interval estimate) that denote the variability and uncertainty in the risk estimate. In the case of the screening-level risk assessment, this can be achieved by using an average, worst, and best case, to illustrate the range of the risk that can be deduced from the available information and the level of certainty that is embedded in the QMRA.

Interval estimates require information about variability and uncertainty. Variability is the result of intrinsic heterogeneity in the input of the risk assessment, such as the variation in *Cryptosporidium* concentration in source water over time, or the variation in the removal of particles by a filtration process over time. Variability can be characterized if sufficient data points are collected. Uncertainty is the result of unknown errors in inputs of the risk assessment, such as errors in the measurement of *Cryptosporidium* or the assumption that certain indicator organisms can be used to describe the removal of *Cryptosporidium* by filtration. Uncertainty can be characterized by specific research activities, for example, to determine the recovery efficiency of the *Cryptosporidium* enumeration method or to compare the removal of *Cryptosporidium* to indicator organisms by filtration.

When sufficient data are available, a probabilistic risk assessment can be performed, where the input is described by statistical distribution functions to describe the confidence interval of the input itself and of the calculated risk.

Good QMRA Practice

Food safety has a longer history of employing microbial risk assessment to facilitate risk management. Several international bodies have produced guidance on good microbial risk assessment practice [13, 72]. The principles of good QMRA practice are also applicable to water safety. General principles are:

- Risk assessment should be clearly separated from risk management.
- Risk assessment should be soundly based on science.
- Risk assessment should be transparent: clear, understandable, and reproducible. It should follow a harmonized procedure based on the accepted standards of best practice.
- The scope and objectives of the risk assessment should be clearly defined and stated at the onset, in collaboration with the risk manager who is going to apply the results.
- The data used are evaluated to determine their quality and relevance to the assessment (taking into account their overall weight in the risk and uncertainty). If data are judged irrelevant or of too low quality, this should be justified. All data that are used are referenced.
- If data are variable, the variability should be documented and taken into account in the risk assessment, preferably in a probabilistic manner.
- All assumptions are documented and explained. Where alternative assumptions could have been made, they can be evaluated together with other uncertainties.
- The risk assessment should include a description of the uncertainties encountered in the risk assessment process. Their relative influence on the risk assessment outcome should be described, preferably in a quantitative (probabilistic) manner. Where point estimates are used for uncertain (or variable) quantities, the selected values should be justified and their influence on the assessment included in the uncertainty analysis.
- Conclusions should reflect the objectives and scope of the risk assessment, and include uncertainties and data gaps.

Uncertainty Analysis

Uncertainty is inherent in risk assessment [54]. Many (if not all) data have a degree of uncertainty. Sources of uncertainty in QMRA include:

- Extrapolation from dose-response data (though, unlike with toxic chemicals, many dose-response data are from human exposure)
- Limitations of pathogen detection methods
- Estimates of exposure

It is important to include the uncertainties in all steps of the risk characterization. The uncertainties in the estimates of exposure are usually dominant. Two approaches are used to determine how the uncertainty in the information in individual steps of the risk assessment affect the uncertainty of the overall risk estimate: sensitivity analysis and Monte Carlo simulation. In sensitivity analysis, the value of each parameter in the risk assessment is varied, one at a time, along the uncertainty range of that parameter (e.g., Average and maximum concentration of a pathogen in water) to determine the effect on the final risk estimate. This procedure generates (1) the range of possible values of the final risk estimate and (2) the uncertainty in which of the parameters contribute most to the uncertainty of the final risk estimate. Sensitivity analysis is typically done in screening-level risk assessments. In probabilistic risk assessments, Monte Carlo simulation is the most widely applied method. Monte Carlo simulation needs a deterministic model for the risk assessment. The uncertainty (and variability) in each of the parameters in the risk assessment is expressed in a probability distribution. The simulation computes a final risk estimate by randomly selecting a value for each parameter in the model from the probability distribution for each parameter. This is repeated many (1,000–10,000) times, each time using a different set of random values from the probability functions. Monte Carlo simulation produces distributions of possible outcome values for the final risk estimate and the shape of the distribution identifies both the general tendency of the risk and the uncertainty of the risk estimate. Also here, the procedure gives information about the contribution of the uncertainty in individual parameters to the uncertainty in the overall risk estimate. While sensitivity analysis evaluates the impact of the uncertainty in each parameter separately and uses few values in the range of possible values of each parameter, Monte Carlo simulation evaluates the impact of the uncertainty in each parameter in combination with all other parameters and uses all possible values and the probability that they occur in the range of each parameter. Burmaster and Anderson [9] published principles of good practice for the use of Monte Carlo simulation in health risk assessments.

Applications of QMRA

The first quantitative microbial risk assessment studies on drinking water were conducted on viruses and *Giardia* [60]. Since the dose-response data from the first human volunteer study on *Cryptosporidium* [12] became available, several authors have performed QMRA for *Cryptosporidium* in water supply (Table 1). This makes the health risk of *Cryptosporidium* through drinking water the most intensively studied object in QMRA studies to date. The overview of QMRA studies for *Cryptosporidium* in water supply illustrates several issues:

1. QMRA studies were conducted to:
   - Evaluate the health risk of *Cryptosporidium* in specific water supply systems or water supply scenarios.
   - Balance the health risk of *Cryptosporidium* in ozonated drinking water to the health risk of bromate formation by ozone [30]. For the assessment of exposure to *Cryptosporidium*, they used raw water monitoring data on *Cryptosporidium*, data on the removal of anaerobic spores by conventional treatment and an ozone disinfection model (the Hom model published by [17]) and a bromate formation model. The ingested dose of oocysts and bromate ions was translated to DALYs to allow comparison of the microbiological and chemical health risk. In their scenario, the health benefits of microorganism inactivation by ozonation outweighed the health losses by bromate formation.
   - Demonstrate the need for additional treatment with UV [1]. They used monitoring data of *Cryptosporidium* in treated water, using a cell-culture-PCR technique to determine the concentration of infectious oocysts in treated water.
   - Demonstrate the need for treatment optimization [46, 48].
   - Illustrate the value of QMRA [47, 48, 59, 66, 68] and relation of QMRA to the Water Safety Plan [49, 65].
   - Evaluate the risk of cryptosporidiosis in different water supply and sanitation scenarios [69].
   - Evaluate the impact of failures in treatment and distribution on the health risk [70]. Failure reports were collected from operational logs/interviews. These failures were translated into an estimate of *Cryptosporidium* (and other pathogen) occurrence (which was the most uncertain step in this QMRA). They indicated that in this system, the health risk associated with normal operation was higher than from the very infrequent and short lasting reported incidents.
   - Prioritize research needs [21], which illustrates how QMRA can be used to determine the relative significance of major, well-controlled and minor, less well-controlled routes of exposure and the impact of moments of reduced treatment performance.
   - Perform a cost-benefit analysis of *Cryptosporidium* regulation that requires additional drinking water treatment for systems with relatively high levels of *Cryptosporidium* in source water [15].
2. Exposure assessment is in many studies hampered by incomplete “site-specific” data. The gaps in the site-specific data are filled by using data from the scientific literature. This is particularly true for the studies in the 1990s. As the use of QMRA progressed, more authors have collected

Microbial Risk Assessment of Pathogens in Water. Table 1 QMRA studies on the risk of *Cryptosporidium* in public water supply

Authors	Exposure assessment	Effect assessment	Outcome	Type	Probability of infection average/95%–range
Medema et al. [47]	*Cryptosporidium* in source water, recovery data [39], viability data [39], removal of oocysts by full scale conventional treatment systems, [39], tap water consumption data [63]	Volunteer study with the Iowa strain [12]	Probability of infection	Probabilistic	3.6×10^{-5} [a] ($3.5 \times 10^{-7} - 1.8 \times 10^{-3}$)
Rose et al. [62]	*Cryptosporidium* in treated water [39]	Volunteer study with the Iowa strain [12]	Probability of infection	Point estimates	5.0×10^{-2} ($4.4 \times 10^{-3} - 1$)
Rose et al. [62]	*Cryptosporidium* in ice prepared from tap water at the time of an outbreak, the latter corrected for the effect of freezing/thawing (90% loss of detectable oocysts) and for the recovery	Volunteer study with the Iowa strain [12]	Probability of infection	Point estimates and comparison of observed and expected illness cases	–
Havelaar et al. [29]	*Cryptosporidium* in source water, recovery data, removal of anaerobic spores by conventional treatment, NL cold tap water consumption data	Volunteer study with the Iowa strain [12]	Probability of infection	Probabilistic	1.3×10^{-4} [a] ($10^{-5} - 10^{-3}$)
Teunis et al. [66]	*Cryptosporidium* in source water, recovery data, viability data [39], removal of anaerobic spores by conventional treatment, NL cold tap water consumption data	Volunteer study with the Iowa strain [12]	Probability of infection	Probabilistic	1.3×10^{-4} [a] ($4 \times 10^{-5} - 4 \times 10^{-4}$)
Teunis and Havelaar [68]	*Cryptosporidium* concentration in source water [5], recovery data [41], viable type morphology [39], removal by storage [66], removal of anaerobic spores by conventional treatment, NL cold tap water consumption data	Volunteer study with the Iowa strain [12]	Probability of infection, illness and DALYs	Probabilistic	No treatment failure: 2.0×10^{-12} 95%: 2.8×10^{-10} Treatment failure: 1.5×10^{-8} 95%: 2.1×10^{-6}

Microbial Risk Assessment of Pathogens in Water. Table 1 (Continued)

Authors	Exposure assessment	Effect assessment	Outcome	Type	Probability of infection average/95%–range
Perz et al. [56]	Assumed concentration of *Cryptosporidium* in tap water, consumption of tap water [63], reduced by 40% for cold tap water consumption and by a further reduction of 33% for AIDS patients	Volunteer study with the Iowa strain [12], assumed threefold higher infectivity for AIDS patients	Probability of infection and illness (probability of illness 0.5 for general population and 1.0 for AIDS patients). Estimated reported cases in general and AIDS population	Point estimates, using two assumed concentrations of *Cryptosporidium* in tap water	$1.0 \times 10^{-3/-2}$ in general population $2.1 \times 10^{-3/-2}$ in AIDS population
Havelaar et al. [30] Gale [20]	*Cryptosporidium* in source water, recovery data, viability data [39], removal of anaerobic spores by conventional treatment, Hom model ozone inactivation [17], NL cold tap water consumption data. The exposure was compared to the exposure to bromate that was formed in the ozonation	Volunteer study with the Iowa strain [12]	DALY	Probabilistic, comparing *Cryptosporidium* to bromate burden of disease	1.0×10^{-3} [a] ($7.6 \times 10^{-4} - 1.5 \times 10^{-3}$)
Haas et al. [24] Haas [26]	*Cryptosporidium* concentration in ice manufactured from tap water during an outbreak, estimation of the inactivation by freezing and thawing, estimation of the duration of the contamination (on onset of cases), attack rate during the outbreak, tap water consumption data [63]	Volunteer study with the Iowa strain [12]	Probability of infection	Point estimate, comparing expected and observed illness	1.1×10^{-2} [b]
Haas et al. [26]	*Cryptosporidium* concentration in distributed water during an outbreak, estimation of the duration of the contamination (on onset of cases), attack rate during the outbreak, assumed 1 L tap water consumption	Volunteer study with the Iowa strain [12]	Probability of infection	Point estimate, comparing expected and observed illness	3.6×10^{-4} [b]

Microbial Risk Assessment of Pathogens in Water. Table 1 (Continued)

Authors	Exposure assessment	Effect assessment	Outcome	Type	Probability of infection average/95%–range
Gale [19, 20]	*Cryptosporidium* in source water [37] and removal of oocysts by full scale conventional treatment systems, [40], data on heterogeneity	Volunteer study with the Iowa strain, including immunity	Probability of infection		1.5×10^{-3} [b]
Haas and Eisenberg [27]	*Cryptosporidium* in different source watersheds, unfiltered system with chlorination, so removal/inactivation by treatment assumed as 0, tap water consumption data [63]	Volunteer study with the Iowa strain [12]	Probability of infection	Point estimate and probabilistic	1.2×10^{-2} 1.2×10^{-3} ($1.2 \times 10^{-4} - 7.7 \times 10^{-2}$)
Medema et al. [48]	*Cryptosporidium* in source water, recovery data, removal of anaerobic spores by conventional treatment, NL cold tap water consumption data	Volunteer study with the Iowa strain [12]	Probability of infection	Point estimate	$1.1 \times 10^{-3} - 3.5 \times 10^{-2}$
	Cryptosporidium in source water, recovery data, removal of bacteriophages by soil passage and of *Cryptosporidium* in soil column studies, NL cold tap water consumption data	Volunteer study with the Iowa strain [12]	Probability of infection	Point estimate	0
	Cryptosporidium in source water, recovery data, viability and genotype data, removal of anaerobic spores by conventional treatment, NL cold tap water consumption data	Volunteer study with the Iowa strain [12]	Probability of infection	Probabilistic	$<1.0 \times 10^{-4}$ with 91% certainty

Microbial Risk Assessment of Pathogens in Water. Table 1 (Continued)

Authors	Exposure assessment	Effect assessment	Outcome	Type	Probability of infection average/95%–range
Westrell et al. [70]	*Cryptosporidium* in source water, removal of particles by conventional treatment, inactivation by disinfection [18, 38], removal of oocysts by membrane filtration [2, 33]	Volunteer study with the Iowa strain [12]	Probability of infection	Probabilistic	Normal operation: 6.0×10^{-4} [a] (6×10^{-6}–4×10^{-2}) Filtration error: 4.0×10^{-5} [a] (6×10^{-7}–2×10^{-3})
	Cryptosporidium in sewage, reports of the water supply on treatment failure and contamination incidents in the distribution network	Volunteer study with the Iowa strain [12]	Probability of infection	Probabilistic	Reservoir contamination: 7×10^{-7} [a] (2×10^{-8}–2×10^{-6})
Masago et al. [46]	*Cryptosporidium* in source water [28], effect of rainfall, viability data [39], failure model for removal by conventional treatment, NL cold tap water consumption data	Volunteer study with the Iowa strain [12]	Probability of infection	Probabilistic	2.0×10^{-4} [a] (2.5×10^{-5} [c] – 2.5×10^{-3})
Gale [21]	Theoretical assumptions in scenario studies of treatment by-pass or failure	Volunteer study with the Iowa strain [12]	Probability of infection	Probabilistic	–
Pouillot et al. [59]	Assumed concentration in distributed water, recovery data, viability data (expert knowledge), French cold tap water consumption	Volunteer study with the Iowa strain for both infection and illness [12], immunodeficient mouse model [75]	Probability of infection and of illness for immunocompetent and immunodeficient persons	Probabilistic	At 2 oocysts/ 100 L: 1.8×10^{-2} 95%: 5.4×10^{-2}
Pouillot et al. [59]	*Cryptosporidium* in distributed water, recovery data, viability data (expert knowledge), French cold tap water consumption	Volunteer study with the Iowa strain for both infection and illness [12], immunodeficient mouse model [75]	Probability of infection and of illness for immunocompetent and immunodeficient persons	Probabilistic	2.1×10^{-2} 95%: 6.7×10^{-2}

Microbial Risk Assessment of Pathogens in Water. Table 1 (Continued)

Authors	Exposure assessment	Effect assessment	Outcome	Type	Probability of infection average/95%–range
Havelaar et al. [30]	*Cryptosporidium* in source water, recovery data, *Cryptosporidium* challenge study of conventional treatment	–	Quality score of exposure assessment factors	Uncertainty analysis	–
Haas et al. [25] JAWWA 88:131	Calculation of a *Cryptosporidium* concentration that corresponds with the 10^{-4} probability of infection (3.27×10^{-5}) oocysts L^{-1} (95% CI: 1.8–6.4 × 10)	Volunteer study with the Iowa strain [12]	Probability of infection	Probabilistic	(1×10^{-4})
Aboytes et al. [1]	*Cryptosporidium* in filtered drinking water, recovery data, infectivity data (cell-culture PCR)	Volunteer studies with the Iowa, UCP and TAMU with Bayesian data-analysis [51]	Probability of infection	Point estimate with confidence interval	8.2×10^{-3} 95%: 1.2×10^{-2}
EPA [15]	*Cryptosporidium* monitoring data (ICR and beyond), recovery data, infectivity fraction, treatment performance credits, USDA consumption data	Volunteer studies with Iowa, TAMU, UCP, using different models	Probability of infection, illness, death and cost	Probabilistic with sensitivity analysis	Scenario evaluation Pre-LT2 filtered: 8×10^{-5} ($<10^{-6}$ – 0.02); unfiltered 0.02 (0.002 to ∼0.5)

[a] Median

[b] Average daily risk of infection during the outbreak

[c] Minimum annual risk

site-specific information about most if not all steps in the exposure assessment.

3. Most studies used the dose-response data of the Iowa strain of *C. parvum* as published by DuPont and coworkers [12]. Over the years, the dose-response relationships of more *C. parvum* strains have been published. One recent study on the risk of *Cryptosporidium* to fire fighters using recycled water used the dose-response data of the TAMU strain of *C. parvum* as this was the most infective strain [11]. Medema [50] present an approach for the use of a *C. parvum* dose-response relation, that combines the dose-response data that are published for four different isolates of *C. parvum* (Iowa, TAMU, UCP and Moredun).
4. The most frequently used health outcome is the probability of infection; a few studies also determined the probability of illness of the general population and the immunodeficient population [56, 59]. Two studies calculated the DALY resulting from the waterborne transmission of *Cryptosporidium* [67, 30].
5. Using the data of the Milwaukee outbreak [44], the calculated probability of infection/illness with QMRA was compared to the observed probability of illness in the outbreak as observed in the epidemiological investigations [24, 26]. The authors concluded that the results of QMRA and epidemiological investigation were consistent. The analysis of the exposure of the Milwaukee residents to *Cryptosporidium* via tap water was hampered by the lack

of timely measurements of *Cryptosporidium* in the contaminated water. Unfortunately, this is the rule rather than the exception in waterborne outbreaks. The concentration had to be inferred from oocyst concentrations found in samples of ice that was prepared at the time of the water supply contamination and was corrected for the expected loss of detectable oocysts after freezing/thawing. The exposure assessment was therefore not very certain. In addition, the reported magnitude of the Milwaukee outbreak has been criticized by [36]. They claim that the background prevalence of gastrointestinal illness in the USA is much higher (1.2–1.4 episodes per person per year, or 0.10–0.12 per person per month) than the prevalence used by [44] (0.005 per person per month). Use of higher background prevalence would drastically reduce the estimated size of the Milwaukee outbreak.

6. The setup of the QMRAs sometimes used point estimates, but more generally a probabilistic approach is used to be able to estimate the level of uncertainty of the calculated probability of infection or illness.
7. Between the different studies, the calculated probability of infection can differ considerably see (Table 1). Within studies, the uncertainty of the risk estimate toward the higher health risk (illustrated by the difference between the average or median risk and the 95% confidence limit) is limited to around a factor of 10.

In general, it can be seen from these examples that QMRA has become an established tool to evaluate health risks of *Cryptosporidium* in (piped) drinking water supplies. QMRA requires input from data on exposure and dose-response and can be done in different levels of complexity. The next paragraphs give examples of the application of QMRA in water and illustrate the stepwise (tiered) approach that can be taken in QMRA and that QMRA can be conducted and be valuable in the absence of site-specific data and in developing countries.

QMRA to Assess the Safety of a Drinking Water Supply

Suppose that a water utility wants to evaluate if its surface water supply is at risk of significantly transmitting *Cryptosporidium* to its consumers, but has no specific information about *Cryptosporidium* in its source water or removal by its water treatment processes. A first exercise to get an idea of the level of risk could be a screening-level QMRA. The information on *Cryptosporidium* levels in source water can be derived from watershed use (see [50]), and for the water treatment processes default log-credits for the removal or inactivation of *Cryptosporidium* are available [49]. For instance, if the water supply system uses a watershed that can be characterized as moderately polluted and treats this source water with off-stream storage reservoirs and a conventional (coagulation/filtration/chlorination) water treatment system, using the scientific database, the expected concentration of *Cryptosporidium* in source water can be estimated at 0.1/L and the removal by the subsequent water treatment processes can be estimated at 0.5 + 2.5 = 3.0 logs removal. Hence, the estimated concentration of *Cryptosporidium* in drinking water is 1×10^{-4}/L. With a conservative best estimate of consumption of cold tap water of 0.78 L/day (3.49 glasses of 0.25 L, [53]), the average probability of exposure to *Cryptosporidium* is 8.7×10^{-5} per person per day. With the combined dose-response relation of the four *C. parvum* strains, the probability of infection is estimated at 3.8×10^{-5} per person per day, which amounts to 1.4×10^{-2} (=1.4%) per person per year. This is a first estimate of the health risk related to *Cryptosporidium* in this specific water supply system. Similarly, such an exercise can be used to evaluate different scenarios of risk management to reduce this risk (if required) such as measures to improve the catchment or install additional treatment processes. An example of a practical application of such a screening-level risk assessment is given in Medema [50], where a large water supply company uses the screening-level QMRA to prioritize risk management of its water supply systems.

Comparing Water Supply Scenarios with QMRA

Piped and non-piped water supply in Uganda [34].

In Kampala, 72% of the population uses piped water supplies. 20% of the population uses piped water through household connections; the rest collects water at standpipes and stores it in-house. The piped water is produced from Lake Victoria water through (coagulation/settling) rapid sand filtration followed by chlorination. The rest of the population (28%) uses protected springs for their water supply.

Data on thermotolerant coliforms were available from Lake Victoria and from the protected springs and the household containers. Using an estimate of the percentage of *E. coli* within the thermotolerant coliforms and an estimate of the percentage of pathogenic *E. coli* within *E. coli*, the thermotolerant coliform concentration data were translated to pathogenic *E. coli* concentrations. For the removal of (pathogenic) *E. coli* by the water treatment processes, the authors used a 3-log credit for the physical removal processes and an additional 2-log credit for the chlorination. This was used to calculate the concentration of pathogenic *E. coli* in drinking water. With data or estimates on consumption of unheated drinking water, dose-response for infection, probability of illness when infected, and disease burden (DALY), the concentration of pathogenic *E. coli* in drinking water was translated into the estimated disease burden by exposure (Table 2).

Similar assessments were made for *Cryptosporidium* and Rotavirus exposure for the population using piped water supply. For *Cryptosporidium*, they showed that treatment failure would result in a very significant increase of the disease burden (from 10^{-4} to 4 DALYs per person per year). The authors have compared the calculated levels of disease burden to the WHO reference level of risk (10^{-6} DALY). Upgrading the treatment would be necessary to achieve this health target, but the authors argue that, given the low level of access to piped water in the home and the disease burden associated with the use of alternative (more contaminated) sources, this would not be cost effective. Improving access to piped water supply in homes, sanitation and hygiene would be more effective in reducing the disease burden.

This example illustrates that QMRA is feasible also in settings with limited data. The authors discuss

Microbial Risk Assessment of Pathogens in Water. Table 2 Assessment of disease burden for pathogenic *E. coli* from different water types (adapted from [34])

	Piped water following treatment	Piped water in distribution	Household storage water	Protected spring water
Raw water quality thermotolerant coliforms/L	150		30	140
Raw water quality *E. coli*/L	143		28.5	133
Raw water pathogenic *E. coli*/L	11.5		2.3	10.6
Treatment effect (log)	5		0	0
Drinking water quality (/L)	1.15×10^{-4}	0.18	2.3	10.6
Consumption of unheated drinking water (L)	1			
Exposure (pathogens/day)	1.15×10^{-4}	0.18	2.3	1.06×10^{1}
Dose-response parameter (exponential)	0.001			
Risk of infection (day)	1.15×10^{-7}	1.80×10^{-4}	2.30×10^{-3}	1.06×10^{-2}
Risk of infection (year)	4.20×10^{-5}	6.57×10^{-2}	8.40×10^{-1}	3.87×10^{0}
Risk of diarrheal disease given infection	0.25			
Risk of diarrheal disease	1.05×10^{-5}	1.64×10^{-2}	2.10×10^{-1}	9.67×10^{-1}
Exposed fraction	0.31	0.1	0.42	0.28
Disease burden (DALYs)	1.04×10^{-6}	5.26×10^{-4}	2.82×10^{-2}	8.67×10^{-2}

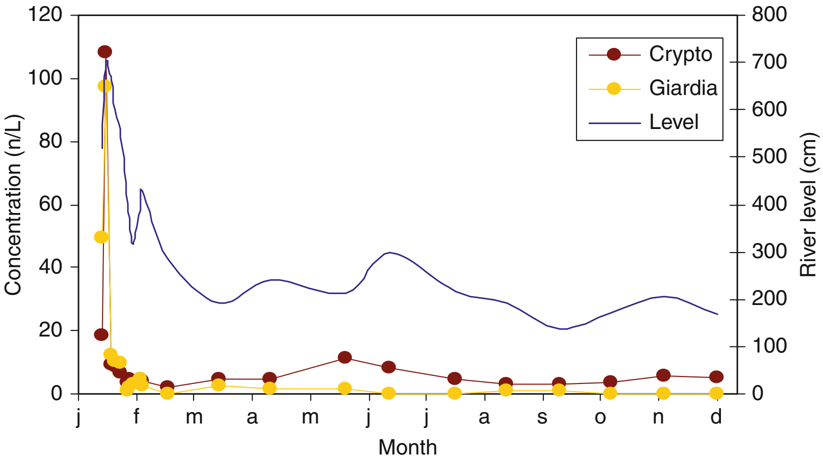
limitations and assumptions used in their study, but illustrate the value of system assessment to inform risk management of the area where control measures will be most effective.

QMRA to Evaluate the Health Risk of Hazardous Events

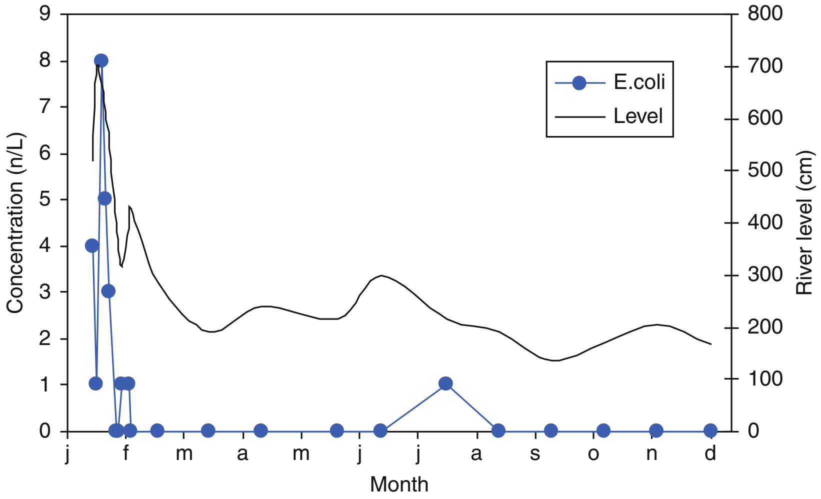
Many outbreaks of intestinal illness caused by consumption of contaminated drinking water in affluent nations have been associated with hazardous events, such as heavy rainfall (both for surface and groundwater systems), failures in a treatment process, failures in the integrity of the infrastructure (wells, distribution network), cross-connections in the distribution network, etc. For an overview, see [35]. Additional hazardous events can be identified for non-piped supplies, especially contamination of the water in storage containers. Also, events that lead to a stop in supply of drinking water (due to power or treatment failure, or indeed absence of sufficient quantities of source water) are hazardous events in themselves, since water is essential for life and hygiene.

Water quality testing can help to identify peak events. Often, peak events can be indicated by simple parameters, such as rainfall, river flow, turbidity, etc., and hence their detection does not require advanced equipment or expertise. It does require knowledge of the water supply system, including its catchment. In Microrisk, a European study on microbial risk assessment of drinking water, information was needed about pathogen occurrence in source (surface) water of the water supply systems under study [49]. Knowing the potential importance of peak events, catchment surveys were conducted to identify contamination sources and to identify events that could lead to peak pathogen contamination of the source water. One system used bank filtration and subsequent treatment to produce drinking water from a large river. Historical (50 years) data on the water level of the river showed that an increase of ≥ 3 m within 5 days occurred 1.1% of the time (3.9 days per year on average). This river level rise was used as a criterion to trigger peak event sampling. A dry weather flow sampling scheme was also in place, with monthly pathogen samples. During monitoring, one peak event was encountered and peak event samples were taken, showing a sharp increase in the concentration of *Cryptosporidium* and *Giardia* concentration in the river (Fig. 3). Event samples were also taken from the bank filtrate. The *E. coli* were detected in the bank filtrate only at the time of the peak event (Fig. 4).

Similarly hazardous events may occur in water treatment (i.e., disinfection failure) or distribution (cross-connection, ingress during main breaks, no pressure period or repair). A QMRA to determine the



Microbial Risk Assessment of Pathogens in Water. Figure 3
Cryptosporidium and *Giardia* in river water during a peak event (Data from [49])



Microbial Risk Assessment of Pathogens in Water. Figure 4

E. coli breakthrough of bank filtration during a peak event in the river see (Fig. 3, data from [49])

health effect of ingress of fecal contamination in municipal piped distribution networks is given in [42]. In the Microrisk project, the health risk associated with several source and treatment hazardous event scenarios in the different water supply systems (called Catchment-to-Tap Systems or CTS) studied was determined and compared to the baseline health risk in these systems in a Monte Carlo simulation [64].

Hazardous events were identified in discussions with local water suppliers and from SCADA data. Of these, five were selected and evaluated with QMRA (Table 3).

In the case of the CTS 1 (a surface water supply) the local managers were concerned about the prospect of a motorway fuel spill and its potential impact on the treatment plant. It was speculated that even small quantities could foul major filters (Rapid Sand Filter and Granular Activated Carbon filters) and reactors (ozone contact tanks) and necessitate cleaning. This led us to simulate a cleanup period of 7 days during which protection was provided by chlorination alone and hence the system was vulnerable to *Cryptosporidium* contamination because of its resistance to chlorine.

It can be seen that the annual risk of infection by *Cryptosporidium* rises by a factor of 1,000 and the estimated probability of infection is much higher than 10^{-4} per person per year. Further, even if the repair period could be reduced to 1–2 days, the additional risk would still be great and hence other action such as a boiled water alert on top of chlorination would need to be considered.

CTS 5 is a surface water supply system with the option to close water intake. If no intake management were in place the average annual risk would have been at least 19 times higher. The impact of a delay in closing the intake was also substantial. This highlighted the need for timely warning of event onset where source extraction is being managed.

CTS 6 included extensive diary and SCADA (Supervisory Control and Data Acquisition) data detailing performance of the chlorination. This information allowed determination whether chlorination failure was occurring. Analysis of the in-line chlorine monitoring data indicated that at worst chlorine dosing failed for a total time of 1.5 h over a 12-month period. The impact of simulated worst-case failure on *Campylobacter* showed a detectable but only small increase in health risk.

The final scenario considered was that of multiple concurrent hazardous events. A concern for CTS 8 and CTS 6 type systems, which draw their supply from a reservoir, is that during high run-off events there can be concurrent polluted input and short-circuiting [32]. Further, storms frequently cause power failures, which could affect treatment plant equipment such as dosing pumps. Two scenarios were considered with

Microbial Risk Assessment of Pathogens in Water. Table 3 Hazardous event impacts on risk

CTS	Pathogen	Hazardous event	Total duration of event	Baseline risk	Baseline + hazardous event risk
				(person^{-1} year^{-1})	
1	*Cryptosporidium*	Loss of filtration due to petroleum spill necessitating cleanup. Only remaining treatment is chlorination	7 days	1.4×10^{-5}	1.7×10^{-2}
5	*Norovirus*	No intake closures leading to periodic high concentration of virus in source water	57 days	$<5.8 \times 10^{-4}$	2.7×10^{-2}
		Delay in intake closure of 4 h for each of 29 events of high virus concentration in source water per year	4.75 days		3.4×10^{-3}
6	*Campylobacter*	Loss of disinfection capacity: total suboptimal chlorination periods based on analysis of SCADA data – worst case of total loss of disinfection assumed	1.5 h	2.5×10^{-6}	3.2×10^{-6}
8	*Campylobacter*	Short-circuiting leads to reduced (1 log) removal in storage reservoir for 24 h. Nine short-circuiting events occur per year	9 days	1.7×10^{-5}	3.4×10^{-5}
		Short-circuiting leads to reduced (1 log) removal in storage reservoir for 24 h. Nine short-circuiting events occur per year. During one of these periods chlorination loss occurs due to power failure for 2.4 h (0.1 days)	0.1 days		1.8×10^{-4}

The risk estimates in brackets are based on upper 95th percentile uncertainty and are derived from upper limit inputs rather than typical source water concentrations

these events in mind. Concurrent contamination of runoff and short-circuiting of the reservoirs were estimated to double health risk for *Campylobacter*. With the combination of a short duration power failure leading to chlorination loss during a storm could increase annualized risk 11-fold in a short time, confirming the need for avoiding or actively managing periods of concurrent hazardous events.

The value of the hazardous event analyses illustrated lies not only in the actual estimates presented. They also demonstrate how QMRA can be used to evaluate events and other hazardous scenarios to produce risk estimates useful for management. In the case of CTS 1, it was clear that filtration shut down even for a short period posed high risks because of the contamination levels in the source water. Selective water intake at CTS 5 is a beneficial management activity. At CTS 6, chlorine dosing was shown to be maintained at a level sufficient to reduce risks arising from plant failure. The CTS 8 analysis showed that baseline operating conditions provide sufficient barrier protection to mitigate a run-off and short-circuiting event, but with a concurrent event (chlorination failure) pose a significant threat.

QMRA for Water Reuse

In (semi) arid conditions, there is (increasing) water scarcity and competition between agriculture and urban uses of this scarce resource. Wastewater is in most cases a reliable (in terms of quantity) source of water and valuable source of nutrients for agriculture. Wastewater reuse in agriculture is a form of water and nutrient recycling that is practiced worldwide, especially in arid and semi-arid areas. Also the (re)use of gray water in urban areas for applications such as toilet flushing in homes, gardening, etc., is becoming more common.

The new WHO Guidelines for safe use of wastewater, excreta, and gray water are based on the Safe Water

Framework (Fig. 1). QMRA is presented in these WHO guidelines as useful tool to estimate the health risks associated with wastewater reuse in different scenarios and for different pathogens. The guidelines contain several references to the application of QMRA in wastewater reuse. In the next paragraphs, three examples of the use of QMRA in water reuse are given.

Comparing Risks Between Different Uses of Reclaimed Wastewater (California)

The first QMRA to estimate the disease risk associated with the reuse of (treated) wastewater was [3]. They evaluated the risk of an infection with enteric viruses (Poliovirus 1 and 3 and Echovirus 12) when chlorinated or unchlorinated tertiary effluent was used for:

- Irrigation of a golf course

  The exposure scenario was a golf course with night time irrigation with tertiary treated wastewater effluent and person golfing twice a week. Each day this person would be exposed to 1 mL of reclaimed water during handling and cleaning of golf balls. The pathogen concentration in this reclaimed water was calculated from data on enteric viruses in chlorinated and unchlorinated effluent and virus decay on the golf field.
- Spray irrigation of food crops

  After spray irrigation, it was assumed that 10 mL of reclaimed water was left on each portion of crops eaten raw. The spray irrigation was stopped 14 days before harvesting and the virus die-off due to desiccation and sunlight exposure was included in the calculation.
- Swimming in recreational water

  This recreational water was assumed to be an impoundment that was, during summer, completely made up out of reclaimed water. No dilution or die-off was assumed. A swimmer was assumed to ingest 100 mL each swimming day and to swim 40 days in a year.
- Groundwater recharge near domestic wells

  This exposure scenario was based on the proposed Californian groundwater recharge regulations. The nearest domestic well was assumed to receive 50% reclaimed water that had been passing through 3 m of unsaturated soil beneath the recharge basin during a period of 6 months. The people drinking from this well were assumed to consume 2 L/day.

  The input data were:

- Concentration of culturable enteric viruses in unchlorinated secondary effluent: 5—734/L (90% and maximum, respectively)
- Concentration of culturable enteric viruses in chlorinated tertiary effluent: 0.01–1.1/L
- Removal of enteric viruses by full tertiary treatment (flocculation, clarification, filtration, chlorination): 5 logs
- Virus decay rate: 0.69/day (first order die-off kinetics)
- Fraction of virus remaining after percolation through the unsaturated soil $c/c_0 = 10^{-0.007\,L}$, where L is the depth of the unsaturated zone in centimeters
- Dose-response parameters for echovirus 12 and poliovirus 1 and 3

The concentration of viruses in reclaimed water was taken from data from surveys of secondary and tertiary effluent. They calculated the exposure to the viruses in the different exposure scenarios. Annual risks were calculated from the maximum concentration found in chlorinated tertiary effluent (1.1 culturable virus unit L^{-1}) and exposure in the different applications (Table 4).

This QMRA showed that the virus risk was highest when reclaimed wastewater was used in recreational

Microbial Risk Assessment of Pathogens in Water. Table 4 Annual risk of exposure to viruses for different applications of reclaimed water

Exposure scenario	Echovirus 12	Poliovirus 1	Poliovirus 3
Irrigation of golf course	1.0×10^{-3}	3.5 E-5	2.5 E-2
Spray irrigation food crops	4.5×10^{-6}	1.5 E-7	1.1 E-4
Recreational impoundment	7.4×10^{-2}	2.6 E-3	8.4 E-1
Groundwater recharge	5.9×10^{-8}	5.4 E-9	2.3 E-8

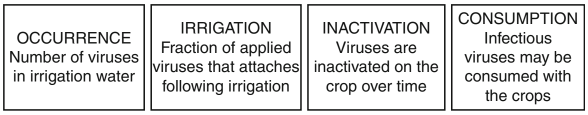
impoundments and golf course irrigation. This maximum concentration was found in only 0.1% of the samples (with 99% of the samples with virus concentrations below the detection limit), so they also calculated the risk with a virus concentration of 1/100 L, which were approximately 100-fold (2 logs) lower.

The value of the QMRA was that it provided a comparative basis for addressing the treatment and fate of enteric viruses in wastewater reuse and showed that the risk can further be mitigated by controlling exposure to reclaimed water.

Health Risk of Reuse for Crop Irrigation (Australia; Probabilistic)

In the previous example, the available data and assumptions were used to generate point estimates. This example shows how the variability in available data can be used to determine the uncertainty that is associated with each of the components in a QMRA.

The use of wastewater for irrigation of food crops that are eaten raw is common practice in many arid and semi-arid regions [57] constructed a QMRA-model for evaluating the risks associated with the consumption of wastewater irrigated lettuce crops. The exposure assessment in this model consisted of four process steps:


OCCURRENCE
Number of viruses in irrigation water

IRRIGATION
Fraction of applied viruses that attaches following irrigation

INACTIVATION
Viruses are inactivated on the crop over time

CONSUMPTION
Infectious viruses may be consumed with the crops

Exposure to viruses was calculated as:

$$\text{Exposure} = N \times f \times S(t) \times q$$

where

N is the number of viruses in the irrigation water applied to the crop

f is the fraction of those viruses that survive the irrigation process and attach to the lettuce plant

$S(t)$ is the fraction of viruses remaining infectious at consumption

q is the quantity of crop consumed

For each step a best estimate and an extreme estimate were selected (Table 5). This allowed analysis of the sensitivity of the QMRA to each of the model parameters.

Microbial Risk Assessment of Pathogens in Water. Table 5 Best and extreme estimates for parameters of exposure to viruses in wastewater reused for irrigation of lettuce

Model component	"Best" estimate	"Extreme" estimate
Virus occurrence	2.6 (virus units L^{-1})	470,000 (virus units L^{-1})
Virus attachment (f)	0.024	0.071
Virus inactivation: S(t) Bi-phasic inactivation $Ct = aC_0 * h1 + (1 - a)C_0h2$	$h1 = 2.5\ day^{-1}$ $h2 = 0.5\ day^{-1}$ a = 0.12%	$h1 = 2.0\ day^{-1}$ $h2 = 0.3\ day^{-1}$ a = 0.96%
Consumption per event q	100 g	300 g

Sources: Californian dataset used by [3, 76]. All other data were derived from [57, 58].

The authors calculated the Factor Sensitivity ($FS = \log(N_{extreme}/N_{best})$, with N being the number of viruses in the extreme or best estimate) for each of the components. Already obvious from the table above is the high impact of the estimate for the virus concentration in wastewater (FS = 5.49). Less obvious from the table is the high impact of the estimate of virus inactivation (FS = 2.2). This is of course time dependent; the authors used 14 days as the time between final irrigation and consumption. A shorter interval reduces the impact of virus inactivation, since the inactivation is less. The uncertainty associated with virus attachment (FS = 0.45) and consumption (FS = 0.48) was considerably less.

This simple mathematical approach yielded not only the risk estimates associated with wastewater reuse for food crop irrigation, but also the (un)certainty associated with each of the components in the exposure of crop consumers to viruses that remain on the crops at the time of consumption.

Guidelines for Safe Reuse (Australia)

QMRA can be used to estimate health risks from exposure to pathogens via wastewater reuse in agriculture, as illustrated in the above examples. In the National guidelines for water recycling in [6], QMRA is used for a different purpose: to calculate health-based performance targets for recycled water systems. In these guidelines, the Australians use a health-based target as a benchmark for safety that has to be met by each water reuse system. They use the health-based target that WHO has defined in their GDWQ: 10^{-6} disability adjusted life years per year (DALY, see Box 1 for more information about this disease burden metric) as their tolerable level of risk.

This health-based target is translated to performance targets for the reuse system with respect to microbial hazards. The concentration of pathogens in the source water for the reuse system (raw/treated sewage, gray water, etc.) and the level of exposure of people to the recycled water (via crops, aerosols, ingestion) determine how much reduction of pathogen exposure is required to meet the 10^{-6} DALY/year target.

In formula:

$$PT = \log(C \times E \times N/DALYd)$$

in which

PT is the performance target (required log reduction)
C is the concentration of pathogens in source water (in these guidelines: 95th percentile of concentration data)
E is the exposure (volume (L))
N is the average frequency of exposure (number/person/year)
DALYd is the pathogen dose that is equivalent to a DALY of 10^{-6} per year, a translation of the 10^{-6} DALY target to a pathogen dose target, taking into account the pathogen's dose-response relation and the fraction of persons that contract illness when infected.

Since sewage and gray water may contain a wide range of pathogens and it is not feasible to do this QMRA for all, it is more practical to select reference pathogens, pathogens that represent a major group of pathogens. The philosophy is that when risk management is aimed at these reference pathogens, the other pathogens from these groups will also be adequately controlled. For protozoa and helminth eggs, *Cryptosporidium* is selected as reference pathogen because it is reasonably infective and more difficult to control by chlorination and filtration than other protozoa or helminth eggs (DALYd is 1.6×10^{-2}, 95th percentile in sewage: 2,000/L). For bacteria, *Campylobacter* is selected because of its infectivity and high prevalence (DALYd is 3.7×10^{-2}, 95th percentile in sewage: 7,000/L). For viruses, rotavirus is selected because of its high infectivity and the availability of dose-response data. Since no data on rotavirus in sewage were available, but data on adenoviruses occurrence were available, these latter data are used and combined with the rotavirus dose-response data (DALYd is 2.5×10^{-3}, 95th percentile in sewage: 8,000/L).

So with concentration C in source water as known and the DALYd as a constant per reference pathogen, the level and frequency of exposure are needed to determine the performance target for the reuse system.

For a range of intended uses of recycled water the associated level and frequency of exposure was (point) estimated from available scientific and statistic data. For example, for exposure by consumption of commercial food crops irrigated with recycled water the level of exposure was estimated at 5 mL for a service of lettuce and 1 mL for a service of other raw produce, with an annual frequency of 70 and 140 services, respectively. Similar exposure estimates were determined for garden irrigation, municipal irrigation, fire fighting, toilet flushing, washing machine use, and cross-connections.

Now the performance target for the use of recycled wastewater for commercial crop irrigation can be calculated:

Exposure for lettuce is 0.005 × 70, for other raw produce 0.001 × 140; this totals to 0.49 L/year

$$\begin{aligned}\mathrm{PT}_{\mathrm{Cryptosporidium}} &= 2,000 \times 0.49/\left(1.6 \times 10^{-2}\right)\\ &= 4.8\ \log\\ \mathrm{PT}_{\mathrm{Campylobacter}} &= 7,000 \times 0.49/\left(3.7 \times 10^{-2}\right)\\ &= 5.0\ \log\\ \mathrm{PT}_{\mathrm{Rotavirus}} &= 8,000 \times 0.49/\left(2.5 \times 10^{-3}\right)\\ &= 6.1\ \log\end{aligned}$$

There are different ways to manage the risk associated with water recycling: prevent pathogens from entering recycled water, remove pathogens from recycled water by treatment processes, and reduce exposure by using restrictions or preventive on-site measures:

restricted access, withholding periods before harvesting, controlled application (drip or subsurface irrigation). The Australian guidelines have assigned default performance credits to a range of treatment processes and on-site preventive measures and give examples of how the combination of these two types of risk management options can be used to achieve safe water recycling.

Box 1. DALY

Disability Adjusted Life Years (DALYs) is as a metric for translating the risk of disease burden a general health burden per case of illness. The DALY accounts for the years lived with a disability (YLD) plus the years of life lost (YLL) due to the hazard (compared to the average expected age of death in a community). One DALY per million people a year roughly equates to one cancer death per 100,000 in a 70-year lifetime [73]. The DALY is calculated as the product of the probability of each illness outcome with a severity factor and the duration (years). Calculation of the DALY contribution per infection is undertaken using:

$$\mathrm{DALY} = \sum_{i=1}^{n} P(\mathrm{ill}|\,\mathrm{inf}) \times P(\mathrm{outcome}_i|\mathrm{ill}) \\ \times \mathrm{Duration}_i \times \mathrm{Severity}_i$$

where n is the total number of outcomes considered

P(ill|inf) is the probability of illness given infection

P(outcome|ill) is the probability of outcome i given illness

Duration$_i$ is the duration (years) of outcome i

Severity$_i$ is the severity weighting for outcome i.

The advantage of using DALYs over an infection risk end point is that it not only reflects the effects of acute end points (e.g., diarrheal illness) but also the likelihood and severity of more serious disease outcomes (e.g., Guillain-Barré syndrome associated with *Campylobacter*). Disease burden per case varies widely, but can be focused on a locality. For example, the disease burden per 1,000 cases of rotavirus diarrhea is 480 DALYs in low-income regions, where child mortality frequently occurs. However, it is only 14 DALYs per 1,000 cases in high-income regions, where hospital facilities are accessible to the great majority of the population. Disease burden estimates for different drinking water contaminants is summarized in Table B1.

Microbial Risk Assessment of Pathogens in Water. Table B1 Summary of disease burden estimates for different drinking water contaminants

	Disease burden per 1,000 cases		
	YLD	YLL	DALY
Cryptosporidium parvum	1.34	0.13	1.47
Campylobacter spp	3.2	1.4	4.6
STEC O157	13.8	40.9	54.7
Rotavirus			
High-income countries	2.0	12	14
Low-income countries	2.2	480	482
Hepatitis-A virus			
High-income countries, 15–49 years	5	250	255
Low-income countries	3	74	77

Source: Reproduced from [31].

Future Directions

The examples given in the previous paragraphs illustrate how QMRA can be applied to assess microbial health risks associated with systems where people may be exposed to pathogens through the use of water. QMRA is used to evaluate individual systems (against health-based targets), compare different systems or scenarios and to evaluate the significance of hazardous events and system failures in municipal piped water supply, but also non-piped water supply, and for wastewater and gray water reuse. Others have also demonstrated the use in recreational waters [4].

Risk assessment also allows comparison of the effort and resources put into the provision of safe water systems and resources allocated to manage other health risks. However, given the current state of the art and especially the lack of available quantitative data, QMRA has to rely partly on assumptions. Given the current level of uncertainty in quantitative risk assessments of water systems, the outcome should be regarded as an indication of the level of safety, rather than an absolute assessment of health risk. The outcome can be used to guide the risk management direction to pathogen control and to select the most appropriate control measures.

The benefit of risk assessment is that it gives a better understanding/breakdown of the problems and of important data. Additionally, the risk concept allows us to focus and prioritize research on the areas where important pieces of information are missing.

Improving the Technique of QMRA

The science of risk assessment is increasingly complex; most of the current QMRA work uses the probability of infection as end point. Infection is the first step in the disease process, but does not reflect the severity of the disease, including potential serious health effects that may arise in a particular subpopulation. Some studies have been using burden-of-disease and cost-of-illness measures [45]. This improves the assessment of the magnitude of the adverse effect of pathogens exposure via water and allows balancing pathogen risks with other risks. The dynamics of infectious diseases with secondary transmission and the effect of immunity and sensitive subpopulations have been largely neglected. Several studies are exploring ways to incorporate these disease dynamics into account [14].

The large variability of pathogens in water and the limited availability of data (especially in relation to peak events) and the variability in treatment efficacy are very important issues to take into consideration in QMRA. More data need to be collected, and monitoring programs of water suppliers should be targeted more toward the provision of information for QMRA. Pathogens to be selected for QMRA should be detectable in the water systems with reliable analytical techniques. The use of reference pathogens, pathogens that are critical for the control measures taken in water supply, is recommended. The variability and limited data available will cause uncertainty in the risk assessment, but compared to chemical risk assessment with large uncertainty factors, this is not inhibitive for the implementation of microbial risk assessment.

Improving the Utility of QMRA

QMRA can be done at different levels of sophistication. Sophisticated QMRA can take considerable amounts of time and resources. The level of detail in the QMRA and the extent of the uncertainty analysis that is needed to address a particular problem has to be appropriate only to the extent that is needed to help risk managers decide. QMRA lends itself well for a tiered approach, where the sophistication increases only if the risk manager requires better information to make a decision.

The National Research Council in the USA has advised USEPA to adopt a framework for risk-based decision making to make risk assessments more useful for risk management decisions [55]. In this framework, improved stakeholder involvement should also help to improve the acceptance and utility of risk assessment.

QMRA is a process that requires input from several disciplines. Researchers that are trained in a specific discipline have to learn to combine their data and knowledge with data and knowledge from other disciplines in a (probabilistic) risk assessment framework. And risk assessment is being extended to address broader questions in environment and health: risk-risk trade-offs and cost-benefit analysis. Development of guidance and training on QMRA is needed to strengthen the capacity of QMRA researchers.

Assessing the microbial risks of water systems is a relatively young field of science. It has the capacity to further professionalize safety management in water by providing science-based, objective, credible and proportionate information to help risk managers make informed decisions.

Bibliography

1. Aboytes R, DiGiovanni G, Abrams F, Rheinecker C, McElroy W, Shaw N, LeChevallier MW (2004) Detection of infectious *Cryptosporidium* in filtered drinking water. J Am Water Works Assoc 96(9):88–98
2. Adham SS, Trussel RR, Gagliardo PF, Olivieri AW (1998) Membranes: a barrier to microorganisms. Water Supply 16(1–2): 336–340
3. Asano T, Leong LYC, Rigby MG, Sakaji RH (1992) Evaluation of the California wastewater reclamation criteria using enteric virus monitoring data. Water Sci Technol 26(7–8):1513–1524
4. Ashbolt NJ, Bruno M (2003) Application and refinement of the WHO risk framework for recreational waters in Sydney, Australia. J Water Health 1(3):125–131
5. Atherholt TB, LeChevallier MW, Norton WD, Rosen JS (1998) Effect of rainfall on *Giardia* and *Cryptosporidium*. J Am Water Works Assoc 90:66–80
6. Australia (2005) Australian guidelines for water recycling: managing health and environmental risks (phase 1). Natural Resource Management Ministerial Council, Environment Protection and Heritage Council, Australian Health Ministers' Conference, Australia

7. Bartram J, Fewtrell L, Stenström TA (2001) Harmonised assessment of risk and risk management for water-related infectious disease: an overview. In: Fewtrell L, Bartram J (eds) Water quality: guidelines, standards and health. IWA Publishing, London, pp 1–16
8. Benford D (2001) Principals of risk assessment of food and drinking water related to human health, ILSI Europe concise monograph series. International Life Science Institute, Belgium, pp 1–43
9. Burmaster DE, Anderson PD (1994) Prnciples of good practice for the use of Monte Carlo techniques in human health and ecological risk assessment. Risk Anal 14(4):477–481
10. CODEX (1997) Codex Alimentarius Commission: procedure manual. Joint FAO/WHO Food Standards Programme. FAO, Rome
11. Deere D, Davison A (2004) Health risk assessment of fire fighting from recycled water. Occasional Paper 11. Water Services of Australia, Sydney
12. DuPont HL, Chappell CL, Sterling CR, Okhuysen PC, Rose JB, Jakubowski W (1995) The infectivity of *Cryptosporidium parvum* in healthy volunteers. N Engl J Med 332(13):855–859
13. EFSA (2009) Transparancy in risk assessment – scientific aspects. Scientific opinion. EFSA J 1051:1–22
14. Eisenberg JNS, Soller JA, Scott J, Eisenberg DM, Colford JM (2004) A dynamic model to assess microbial health risks associated with beneficial uses of biosolids. Risk Anal 24(1):221–236
15. EPA (Environmental Protection Agency) (2005) Economic analysis for the final long term 2 enhanced surface water treatment rule. EPA, Office of Water report EPA 815-R-06-001. EPA, Washington
16. Fewtrell L, Bartram J (2001) Water quality: guidelines, standards and health. Risk assessment and management for water related infectious diseases. IWA Publishing, London
17. Finch GR, Black EK, Gyurek L, Belosevic M (1993) Ozone inactivation of *Cryptosporidium parvum* in demand-free phosphate buffer determined by in vitro excystation and animal infectivity. Appl Environ Microbiol 59(12):4203–4210
18. Finch GR, Gyurek LL, Liyanage LRJ, Belosevic M (1997) Effect of various disinfecion methods on the inactivation of *Cryptosporidium*. American Water Works Association Research Foundation, Denver
19. Gale P (1996) Developments in microbiological risk assessment models for drinking water – a short review. J Appl Bacteriol 81:403–410
20. Gale P (2000) Risk assessment model for a waterborne outbreak of cryptosporidiosis. Water Sci Technol 41(7):1–7
21. Gale P (2002) Using risk assessment to identify future research requirements. J Am Water Works Assoc 94(9):30–42
22. Gale P, Stanfield G (2000) *Cryptosporidium* during a simulated outbreak. J Am Water Works Assoc 92(9):105
23. Haas CN (1997) Importance of distributional form in characterizing inputs to Monte Carlo risk assessments. Risk Anal 17(1):107–113
24. Haas CN (2000) Epidemiology, Microbiology, and Risk Assessment of Waterborne Pathogens Including *Cryptosporidium*. J Food Prot 63(6):827–831
25. Haas CN, Crockett CS, Rose JB, Gerba CP, Fazil AM (1996) Assessing the Risk Posed by Oocysts in Drinking Water. J Am Water Works Assoc 88(9):131–136
26. Haas CN, Rose JB, Gerba CP (1999) Quantitative microbial risk assessment. Wiley, New York
27. Haas CN, Eisenberg JNS (2001) Risk assessment. In: Fewtrell L, Bartram J (eds) Water quality: guidelines, standards and health. IWA Publishing, London, pp 161–183
28. Hashimoto A, Hirata T (1999) Occurrence of *Cryptosporidium* oocysts and *Giardia* cysts in Sagami river, Japan. In: Proceedings of the Asian water quallity 1999; 7th reginal IAWQ conference. Taipei, Taiwan, vol 2, pp 956–961
29. Havelaar AH, Teunis PFM, Medema GJ (1996) Risk assessment of waterborne pathogens. Presented at the symposium waterborne pathogens, Bonn, 22–25 May
30. Havelaar AH, den Hollander AEM, Teunis PFM, Evers EG, van Kranen HJ, Versteegh JFM, van Koten JEM, Slob W (2000) Balancing the risks and benefits of drinking water disinfection: disability adjusted-life years on the scale. Environ Health Perspect 108(4):315–321
31. Havelaar AH, Melse JM (2003) Quantifying health risks in the WHO guidelines for drinking water quality. A burden of disease approach. Report 734301022, RIVM, Bilthoven
32. Hipsey MR, Brookes JD, Antenucci JP, Burch MD, Regel R (2004) A three-dimensional model of Cryptosporidium dynamics in lakes and reservoirs: a new model for risk management. Int J River Basin Manag 2(3):181–197
33. Hirata T, Hashimoto A (1998) Experimental assessment of the efficacy of microfiltration and ultrafiltration for *Cryptosporidium* removal. Water Sci Technol 41(7):103–107
34. Howard G, Pedley S, Tibatemwa S (2006) Quantitative microbial risk assessment to estimate health risks attributable to water supply: can the technique be applied in developing countries with limited data? J Water Health 4(1):49–66
35. Hrudey S, Hrudey E (2004) Safe drinking water, Lessons from recent outbreaks in affluent nations. IWA Publishing, London, 486 pp
36. Hunter PR, Syed Q (2001) Community surveys of self-reported diarrhoea can dramatically overestimate the size of outbreaks of waterborne cryptosporidiosis. Water Sci Technol 43(12): 27–30
37. Hutton P, Ashbolt N, Vesey G, Walker J, Ongerth J (1995) Cryptosporidium and Giardia in the aquatic environment of Sydney, Australia. In: Betts WB et al (eds) Protozoan parasites and water. Royal Society of Chemistry, Cambridge
38. Korich DG, Mead JR, Madore MS, Sinclair NA, Sterling CR (1990) Effects of ozone, chlorine dioxide, chlorine, and monochloramine on *Cryptosporidium parvum* oocyst viability. Appl Environ Microbiol 56(5):1423–1428
39. LeChevallier MW, Norton WD, Lee RG (1991) Occurrence of *Giardia* and *Cryptosporidium* spp. in surface water supplies. Appl Environ Microbiol 56:2610–2616
40. LeChevallier MW, Norton WD (1995) *Giardia* and *Cryptosporidium* in raw and finished water. J Am Water Works Assoc 87(9):54–68

41. LeChevalier MW, Norton W, Abbaszadegan M, Atherholt T, Rosen J (1998) Development of a monitoring strategy to determine variations in *Giardia* and *Cryptosporidium* levels in a watershed. In: Proceeding of source water protection international 1998, Dallas
42. van Lieverloo JHML, Blokker EJ, Medema GJ (2007) Quantitative microbial risk assessment of distributed drinking water using faecal indicator incidence and concentrations. J Water Health 5(Suppl 1):131–149
43. Macler BA, Regli S (1993) Use of microbial risk assessment in setting United-States drinking water standards. Int J Food Microbiol 18:245–256
44. MacKenzie WR, Hoxie NJ, Proctor ME, Gradus MS, Blair KA, Peterson DE, Kazmierczak JJ, Addiss DG, Fox KR, Rose JB, David JP (1994) A massive outbreak in Milwaukee of *Cryptosporidium* infection transmitted through the public water supply. N Engl J Med 331(3):161–167
45. Mangen MJJ, Havelaar AH, Bernsen RAJAM, Van Koningsveld R, De Wit GA (2005) The costs of human Campylobacter infections and sequelae in the Netherlands: a DALY and cost-of-illness approach. Food Econ Acta Agric Scand 2(1):35–51
46. Masago Y, Katayama H, Hashimoto A, Hirata T, Ohgaki S (2002) Assessment of risk of infection due to *Cryptosporidium parvum* in drinking water. Water Sci Technol 46(11/12):319–324
47. Medema GJ, Teunis PFM, Gornik V, Havelaar AH, Exner M (1995) Estimation of the *Cryptosporidium* infection risk via drinking water. In: Betts WB et al (eds) Protozoan parasites and water. Royal Society of Chemistry, Cambridge, pp 53–56
48. Medema GJ, Hoogenboezem W, van der Veer AJ, Ketelaars HAM, Hijnen WAM, Nobel PJ (2003) Quantitative risk assessment of Cryptosporidium in surface water treatment. Water Sci Technol 47:241–247
49. Medema GJ, Loret JF, Stenström TA, Ashbolt NJ (2006) Quantitative microbial risk assessment in the water safety plan. Report Microrisk, Kiwa Water Research, Nieuwegein. www.microrisk.com
50. Medema GJ (2009) Risk assessment of *Cryptosporidium* in drinking water. World Health Organization, Geneva, Switzerland
51. Messner MJ, Chappell CL, Okhuysen PC (2001) Risk assessment for *Cryptosporidium*: a hierarchical Bayesian analysis of human dose response data. Water Res 35(16):3934–3940
52. Michaels D (2008) Doubt is their product: how industry's assault on science threathens your health. Oxford University Press, New York
53. Mons MN, van der Wielen JHML, Blokker EJM, Sinclair MI, Hulshof KFAM, Dangendorf F, Hunter PR, Medema GJ (2007) Estimation of the consumption of cold tap water for microbiological risk assessment: an overview of studies and statistical analysis of data. J Water Health 5(Suppl 1):151–170
54. Morgan MG, Henrion M (1990) Uncertainty. A guide to dealing with uncertainty in quantitative risk and policy analysis. Cambridge University Press, New York, 332 p
55. NRC (2009) Science and decisions, Advancing risk assessment. National Research Council. The national Academies Press, Washington, DC
56. Perz JF, Ennever FK, Le Blancq SM (1998) Cryptosporidium in tap water: comparison of predicted risks with observed levels of disease. Am J Epidemiol 147(3):289–301
57. Petterson SR, Ashbolt NJ, Sharma A (2001) Microbial risks from wastewater irrigation of salad crops: a screening-level risk assessment. Water Environ Res 73(6):667–672
58. Petterson SA, Ashbolt NJ (2005) WHO guidelines for the safe use of wastewater and excreta in agriculture. Microbial Risk Assessment Section. www.who.int/wsh
59. Pouillot R, Beaudeau P, Denis J-B, Derouin F (2004) A quantitative risk assessment of waterborne cryptosporidiosis in France using second-order Monte Carlo simulation. Risk Anal 24(1):1–17
60. Regli S, Rose JB, Haas CN, Gerba CP (1991) Modeling the risk from Giardia and viruses in drinking water. J Am Water Works Assoc 83:76–84
61. Regli S, Odom R, Cromwell J, Lustic M, Blank V (1999) Benefits and costs of the IESWTR. J Am Water Works Assoc 91(4): 148–158
62. Rose JB, Lisle JT, Haas CN (1995) Risk assessment methods for *Cryptosporidium* and *Giardia* in contaminated water. In: Betts WB et al (eds) Protozoan parasites and water. Royal Society of Chemistry, Cambridge, pp 238–242
63. Roseberry AM, Burmaster DE (1992) Lognormal distributions for water intake by children and adults. Risk Anal 12:99–104
64. Roser D, Petterson S, Signor R, Ashbolt N, Nilsson P, Thorwaldsdotter R (2006) How to implement QMRA to estimate baseline and hazardous event risks with management end uses in mind. Report Microrisk, Kiwa Water Research, Nieuwegein. www.microrisk.com
65. Smeets P (2008) Stochastic modelling of drinking water treatment in quantitative microbiological risk assessment. PhD thesis, Delft University, The Netherlands
66. Teunis PFM, Medema GJ, Kruidenier L, Havelaar AH (1997) Assessment of the risk of infection by *Cryptosporidium* and *Giardia* in drinking water from a surface water source. Water Res 31(6):1333–1346
67. Teunis PFM, Havelaar AH (1997) *Cryptosporidium* in drinking water. Evaluation of the ILSI/RSI quantitative risk assessment framework. RIVM, Bilthoven, The Netherlands
68. Teunis PFM, Havelaar AH (1999) *Cryptosporidium* in drinking water. Evaluation of the ILSI/RSI quantitative risk assessment framework. Report no. 284 550 006 RIVM, Bilthoven
69. Westrell T (2004) Microbial risk assessment and its implications for risk management in urban water systems. Thesis Linköping University, Linköping
70. Westrell T, Bergstedt O, Stenstrom TA, Ashbolt NJ (2003) A theroretical approach to assess microbial risks due to failures in drinking water systems. Int J Environ Health Res 13:181–197
71. WHO (1993) Guidelines for drinking water quality, 2nd edn. WHO, Geneva
72. WHO (1999) Principles and guidelines for the conduct of microbiological risk assessment. WHO, Geneva, CAC/GL-30
73. WHO (2004) Guidelines for drinking water quality, 3rd edn. WHO, Geneva

74. WHO (2006) Guidelines for the safe use of wastewater, excreta and grey water. WHO, Geneva
75. Yang SS, Benson SK, Du C, Healey MC (2000) Infection of immunosuppressed C57BL/6N adult mice with a single oocyst of *Cryptosporidium parvum*. J Parasitol 86:884–887
76. Yates MV, Gerba CP (1998) Microbial considerations in wastewater reclamation and reuse. In: Asano T (ed) wastewater reclamation and reuse, vol 10, Water Quality Management Library. Technomic Publishing, Lancaster, pp 437–488

Mining and Its Environmental Impacts

Jörg Matschullat, Jens Gutzmer
TU Bergakademie Freiberg, Institute of Mineralogy
and Helmholtz Institute Freiberg for Resource
Technology, Freiberg, Saxony, Germany

Article Outline

Glossary

Biota All life forms.

Decommissioning Removal of something from active status.

Eco-efficiency analysis Analysis of realizing the concept of creating goods and services with fewer resources and less waste and pollution.

Exploitation Act of using something (mineral resources) for any purpose.

Exploration Process of finding mineral resources for the purpose of mining.

Karst A geological feature in relatively soluble rocks, e.g., limestone, where sinkholes, caves, and similar hollows are formed above and below ground.

Lithosphere The outer rocky shell of planet Earth, comprising the oceanic and continental crust and part of the upper Earth mantle.

Long-term effect A change that will last or have an influence over a long period of time.

Nachhaltigkeit German for “sustainability,” first used in 1713 in Germany.

Open-pit excavation Process of extracting minerals from surface deposits.

Recultivation Making raw mineral soils (brownfields) fertile again through bioengineering and refertilization.

Rehabilitation Restoring land after some process has damaged it.

Remediation Removal of pollution or contaminants from the environment.

Sinkhole A natural depression at the Earth surface generated by subsurface erosion, particularly in karst areas.

Slag Partially glassy by-product from smelting ore, mostly consisting of a silicate matrix with metal oxides.

Sustainable mining Mining method that does not compromise environmental quality.

Definition of the Subject

The environmental impact of mining is the influence that mining activities have on the natural conditions and world in which humans and all biota live. The impact may involve diverse forms of environmental change or damage, from short- to long-term effects and from highly spatially restricted to long-distance consequences.

Just as any kind of human activity, mining has an inherent and partly unavoidable impact on the environment. From the first steps of exploration via exploitation and (ore) processing to the final stages of decommissioning and rehabilitation, environmental hazards and risks may be encountered and need to be addressed. The potential impacts and long-term aftermath of mining operations are manifold. Whether in fact and to which extent the impacts do lead to detrimental consequences in any one of the environmental compartments (atmosphere, hydrosphere, pedosphere, biosphere, cryosphere, and lithosphere) is difficult to predict. A thorough investigation of local conditions – both boundary conditions and operation-related conditions – is needed to answer that crucial question. Modern mines can be operated in a manner much less detrimental than was the standard up until only recently. Parallel to the mining industry’s awareness of issues connected with the environment, however, new challenges appear: The exploitation of lower concentrations of the valuable constituents (e.g., minerals or metals) presents most demanding challenges that deal with difficult conditions and involve a larger footprint of mining and more complex approaches to beneficiation. Increasing challenges also apply to marine mining, be it nearshore or offshore; be it for diamonds and other placer deposits, oil, and gas; or be it for manganese nodules that are being procured in increasingly deeper marine environments [1].

Developing and implementing sustainable mining practices are tasks that have been high on the agenda of the international mining industry. These need to become the global standard to curtail the most long-lasting and detrimental impacts of mining. Although the necessary knowledge base is rapidly becoming available; there is still a need for basic research to further establish and foster sustainable solutions. Due to its exotic nature and less likelihood to disturb Earth's immediate biosphere and equilibrium, "extraterrestrial mining" is not dealt with here.

Introduction: Sustainable Mining – An Oxymoron?

Environmental impacts of mining appear to be most well known all over the world – almost beyond the necessity of further elucidation and questioning ([2, 3]; Table 1). The mining industry has recognized its impact on the environment and has identified the control and restriction of such impact as one of its key challenges [4, 47])

From a workable standpoint, mining encompasses a very large array of activities. Especially in highly industrialized nations, there is major resistance of societies against mining activities. For the building and construction industry, even aggregate materials from quarries, and sand and gravel from open-pit excavations are increasingly under scrutiny in densely populated areas. Generally, much larger open-pit or underground mines for rock salt, metalliferous ores, or precious stones are often no longer perceived as indicators of economic well-being and development, but rather symptomatic of visual, acoustic, and environmental perturbations with detrimental impact. The same is certainly true for energy resources, namely gas, oil, coal, tar sands, and uranium ores, whether or not these are being mined on land or in marine shelf areas. Not only the general public, but a considerable part of the decision makers in both industry and politics (at least in the western world) has developed a stance that mining is per se a dirty business and that its related activities can be left to (mostly) developing countries.

In the late twentieth century, the idea that "industry and the developed nations would always be able to buy the necessary commodities" prevailed. Since the advent of the twenty-first century, this position has been increasingly under scrutiny, simply because the growing world population demands increasing amounts of raw materials. To find acceptance and support in society, any future mining activity will demand a state-of-the-art environmental management and has to contribute to sustainable development [5]. Furthermore, the role of the mining industry is set to increase, as technological advancement demands rapidly increasing supplies of a rising number of raw materials that have never before found a significant industrial application, e.g., rare earth elements (REE) or lithium (Li).

To clarify current environmental issues connected to mining and to be able to develop alternatives to practices that are currently widely used, an understanding of the history of mining is needed, as well as an overview of the environmental effects of mining, differentiated by its relevant phases: exploration, exploitation and processing, decommissioning, and rehabilitation. Thereby, potentially negative impacts may be largely avoided or significantly abated if intelligent and foresighted precaution is taken. At the end of this contribution, future perspectives and the pathway to sustainable mining shall be evaluated.

A Concise Review of Mining History

Mining has been with mankind for much more than 40,000 years already (Paleolithic), when commodities were procured from surface and even underground deposits from various places on several continents, in order to obtain flint stones for axes and arrowheads, clay and loam for pottery and construction (e.g., [6]), or iron oxide (hematite) for cosmetic purposes [7, 8]. The interest in metals developed later, as the related human activities in metal working became so widespread that entire epochs were named accordingly (Bronze Age, Iron Age, etc.) [9, 10]. With the advent

Mining and Its Environmental Impacts. Table 1 Potential (and real) environmental impact of mining on environmental compartments

Compartment	Potential environmental impact and spatial extent
Atmosphere	Release of (toxic) gases (e.g., SO_2 emissions from sulfide ore roasting, CO_2 and CFC release from aluminum processing), dusts, and aerosols: very-short- (local) to long-range transport may contaminate vegetation cover and other biota, soil, and water. Often, the burning of fossil fuels (e.g., from energy generation) has more detrimental effects than the mining operation itself
	Alteration of local air humidity: effects on local microclimate and thus biota
Hydrosphere	*Surface water*: water level fluctuations, water losses, floods, direct contamination; accidental connection of surface water and soluble ore deposits with local to small regional effects
	Groundwater: direct contamination via seepage water from aboveground or directly from within the mine operation; lowering (due to new permeabilities) or rising of water table (due to ground softening and compression), vertical fluctuations of the water table, causing local underground erosion and loss of rock stability, hydraulic filling of underground cavities and aquifers after stopping the water pumping at mine closure – affects the aquifer extension
	Drainage and seepage water: saline water, acid water, and alkaline water, each with specific toxins – affects surface and groundwaters
	Coastal and marine waters: direct pollution by spills; placer deposit mining disrupts beach systems; use of deep-sea deposits threatens (rare) marine life; mostly local effects
Pedosphere	Soil loss (open-pit and underground operations) – competition for land use; large volumes of waste-rock heaps and tailings deposits; soil contamination (and water) by water spills and seepage of contaminated waters from slag and waste heaps, tailings deposits, improper operation, etc. – local effects
Biosphere	Disturbance of ecosystems, disruption of food chains, eviction of (key) species; silicosis and in general, inhalation of fibers (asbestos mining) as a health hazard to workers and high ambient dust concentrations in the vicinity of operations – local to small regional effects
Lithosphere	Mine structure (surface: collapse, subsidence; underground: pillar breaking, slab breaking), aquifer mine operation: surface overloading, surface vibrations and shaking (blasting), mine sludge, mine tailings, slag heaps – local effects
Anthroposphere	Damage to infrastructure (transport, buildings, etc.) due to surface movements and subsidence – local effects

of even more sophisticated technologies in the Chinese and Roman Empires, the spectrum of sought-after elements had expanded and included components such as silver (Ag), arsenic (As), gold (Au), copper (Cu), iron (Fe), mercury (Hg), lead (Pb), and tin (Sn) minerals [11–13]. There is no doubt that mining has been pursued by man millennia before the Industrial Revolution and on all continents (except for Antarctica), and independent of the global European influences that started with their conquest in the fifteenth century [14]. Following the Industrial Revolution in the late eighteenth to mid-nineteenth centuries, this range of elements not only increased, but new orders of magnitude were reached in the demand for metals and other commodities, including fast-increasing amounts of nonrenewable energy resources (coal, oil, and gas – in that sequence). Today, the timing of Peak Oil is discussed in parallel and just as intensively as the possible shortage of REE and other metals that our modern industrial society technologically and economically depend upon. At the same time, possible causes for the collapse of historical civilizations are examined, and there is evidence that mining may have contributed to such human self-destruction in the past [15].

It was largely not before post-WWII economic recovery that people in the involved industrial countries started to look at the environmental impact of these mining activities. Using European history, the cradle of the Industrial Revolution, as an example, voices from almost two millennia back deserve mentioning.

Central Europe was mainly forest-covered until the advent of the Early Middle Ages. Mining had remained a rather small-scale business, even though during Roman times (ca. 300 AD), comparatively sophisticated mining and smelting technology is already known from various places (e.g., Harz Mountains, Germany [11]). The push towards the eastern frontier by Gallic and Germanic people under the guidance of Charlemagne (around AD 800) led to new settlements, a subsequent period of significant forest removal and a new period of mining exploration and exploitation. Cities like Annaberg and Freiberg, Erzgebirge in Saxony, Germany; Kutna Hora in Bohemia (Czech Republic); and many others may serve as examples. There, silver (Ag) was found, a key resource for coin making and luxury items – much like today. These successful mining cities developed, and production increased. Around the 1500s, most of the higher-elevation forests had been cut down already, leaving vast stretches of almost tree-barren landscape – much of which was used for agriculture.

In 1557, Georgius Agricola (Latinized version of his name Georg Bauer), a German medical doctor and early allround scientist from Chemnitz in Saxony, published the first and most comprehensive book (12 volumes) on mining and its implications, "De re metallica" (about the metal issues) [46]. His book did not only describe mineral exploration techniques (even touching the use of metallophilic plants), exploitation, and smelting techniques, but also explicitly introduced the reader to the detrimental side effects of mining. He reviewed the "bad smokes" and their effects on biota; he described the barren land where no plants would want to grow and the "dead waters" where fish would no longer live or spawn [16]. Indeed, one has to imagine such metalliferous provinces in Europe as being mainly forest-free areas after a few centuries of steadily increasing mining and smelting activities. The wood was needed and used both to fuel processes and to build support structures and equipment (water wheels, water ducts, etc.) in the mines.

As of the eighteenth century, such evidence on the effects of mining became even more prominent, and more publications related to these issues emerged. In 1713, the first book on sustainability was published to introduce the concept and coin the term "sustainability" (in German *Nachhaltigkeit*) [17, 18]). Hannß Carl von Carlowitz, a Freiberg mining engineer, was also responsible for the wood supply for the local mines and noticed the increasing depletion of this valuable natural resource. Von Carlowitz wrote that a sustainable forest management was then urgently necessary to avoid (and repair) the damage resulting from mining and smelting activities if this business was to continue. As a matter of fact, those days saw a significant decline in mine productivity for various reasons: an increasing lack of wood, a steadily increasing demand for more sophisticated technologies, and permanent wars between the small states and provinces in Central Europe. Until those days, mining and smelting were done as many centuries before, with minor technological advances. This now changed rather rapidly, with first the introduction of stipends for gifted young men (non-aristocrats) to receive a higher education in mining and in 1764 with the foundation of the world's first mining academy (then named *Bergakademie*), today known as the Technical University (TU) Bergakademie Freiberg.

Soon, new technologies were introduced that increased the efficiency of mines by reducing the water and wood (fuel) demand per volume of ore. At the same time, forestry was developed as a scientific field, and in 1811–1816, the world's first forest academy was founded in Tharandt, Saxony. Ever since, it serves to educate future foresters and forest scientists and is today part of the Technical University Dresden. Today, the Erzgebirge is largely forested again (i.e., similar to the Harz Mountains and the Black Forest in Germany, and many other historical mining areas) with a forest cover of about 30%.

Environmental Impacts of Mining

Throughout many historical mining districts, the less noticeable centuries-old legacy of mining is still perceptible to the trained eye. It yields many helpful lessons on avoiding further environmental damage and developing sustainable mining techniques.

In general, the environmental impact of mining takes place on many levels and may affect most environmental compartments – atmosphere, hydrosphere, pedosphere, biosphere, and lithosphere, and under certain aspects even the cryosphere. Some of the key “priority pollutants” are metals that are being liberated through mining and related activities (Table 2).

Mining requires exploration to identify the exploitation potential of a mineral deposit. Related investigations may include not only geophysical (electric, electromagnetic, gravity, and seismic investigations) and geochemical work at the surface (digging of pits, trenches, or rock cuts) but also drilling activities to verify obtained results. This enables and supports 3D modeling of the ore body, a basis for reducing technical and financial risks. Following a successful exploration phase, and depending on the decision for aboveground (open-pit) or underground mining, the required mining infrastructure will be developed and ensued by rather large surface excavation or the construction of shafts and tunnels. In most cases, extensive aboveground facilities are built concurrently, which encompass infrastructure for processing, workshops, storage, and a general infrastructure of offices, remote control rooms, transport access from helicopter ports and airfields, road and train access to the electrical and water

Mining and Its Environmental Impacts. Table 2 Priority pollutants: metals from natural and mining-related sources[a]

Element	Natural source	Anthropogenic source	Common forms in waste
Ag	Native metal (Ag), chlorargyrite (AgCl), acanthite (Ag_2S); Cu, Pb, Zn ores	Mining	Metallic Ag, Ag–CN complexes, Ag halides, Ag thiosulfates
As	Metal arsenides and arsenates, complex sulfide ores (arsenopyrite, FeAsS), arsenolite (As_2O_3), volcanic gases, geothermal springs	Pyrometallurgical industry, soil heaps and tailings, smelting, mine drainage	As oxides (oxyanions), organo-metallic forms, methylarsinic acid ($H_2AsO_3CH_3$), dimethylarsinic acid ($(CH_3)_2AsO_2H$)
Cd	Zn sulfide ores	Mining and smelting, mine drainage	Cd^{2+} ion, Cd halides and oxides, Cd–CN complexes, $Cd(OH)_2$ sludges
Cr	Chromite ($FeCr_2O_4$)	Pyrometallurgical industry	Metallic Cr, Cr oxides (oxyanions), Cr^{3+} complexes with organic and inorganic ligands
Cu	Native metal (Cu), chalcocite (Cu_2S), chalcopyrite ($CuFeS_2$), bornite (Cu_5FeS_4)	Mining and smelting, pyrometallurgical industry, mine drainage	Metallic Cu, Cu oxides, Cu–humic complexes, alloys, Cu ions
Hg	Native metal (Hg), cinnabar (HgS), degassing from Earth’s crust and oceans	Mining and smelting, mine drainage	Organo–Hg complexes, Hg halides and oxides, Hg^{2+}, $(Hg_2)^{2+}$, Hg^0
Ni	Pentlandite ($(Fe,Ni)_9S_8$), Ni hydroxy-silicate minerals	Mining and smelting	Metallic Ni, Ni^{2+} ions, Ni amines, alloys
Pb	Galena (PbS)	Mining and smelting, mine drainage	Metallic Pb, Pb oxides and carbonates, Pb-metal–oxyanion complexes
Sb	Stibnite (Sb_2S_3), geothermal springs	Pyrometallurgical industry, smelting, mine drainage	Sb^{3+} ions, Sb oxides and halides
Se	Polymetallic base metal sulfide ores	Smelting	Se oxides (oxyanions), Se–organic complexes
Tl	Polymetallic base metal sulfide ores	Pyrometallurgical industry	Tl halides, Tl–CN complexes
Zn	Sphalerite (ZnS)	Mining and smelting, pyrometallurgical industry, mine drainage	Metallic Zn, Zn^{2+} ions, Zn oxides and carbonates, alloys

[a] Table modified and focused on mining-related activities after Adriano [49] and Sparks [10]

supply, ore dressing and smelting facilities, and room for waste rocks and tailings deposits. Any one of these units must be seen as an integral part of the mining activity, each with a potential environmental imprint.

Exploration Phase

In the exploration phase, already and depending on the previous land use, land has to be cleared and roads and (minor) infrastructure constructed. Climatological and local conditions define the intensity and duration of exploration activities and thus play a role in the environmental impact. Largely, exhaust fumes and dust emissions may influence air quality during this operation [19]. In general, such works and the related noise emissions have a highly restricted local impact that will stop or rapidly decrease with the end of the exploration activities. Water resources can be impacted during exploration activities by improper handling of equipment and insufficient control of exploration drilling (spillage of drilling additives, oil losses, etc.). Primarily, temporary losses in aquatic biodiversity result; hence, the shorter the operation, the easier is the recovery. Yet, related impacts may remain evident for years and even decades. Soils have a much longer "memory" for human activities. The construction of drilling platforms (pressure and surface sealing) and the drill waste materials (including potentially toxic matter) may leave imprints for many decades or even centuries (arctic environments = potential impact on the cryosphere), albeit again, on a very local scale. Biospheric impact may be of critical importance since it is directly related to all other environmental compartments. Here, environmental impact assessment studies may be helpful prior to starting with the mining phase. Such assessments do not necessarily impede the progress of the exploration project and principally depend on the available ecosystem or site-specific knowledge of biologists or ecologists. These evaluations are in most cases restricted in time, and recovery is possible, provided that state-of-the-art operations and precaution are applied. The crucial and well-known risks related to immediate accidents (fatalities and injuries) and health problems during the mining process itself are not dealt with here. Impact on the lithosphere is restricted to excavations and boreholes themselves and may pose challenges mainly in unstable surface and in karst environments, e.g., triggering unwanted water pathways or rock-mechanical instabilities. In general, and particularly at locations with unsuccessful exploration activities, related legacies of failed prospecting and exploration may impact future land use much later due to non-documented activities in the mining phase that may compromise the free choice of subsequent land use.

While most environmental impacts are small-scale and short-term in the exploration phase, incognizant or careless practices can lead to serious consequences. Therefore, before start-up (and beyond closure) of each mining operation, responsible exploration activities should include a priori environmental assessment studies [20]. Related important work that involves the post-mining operations, such as reclamation and rehabilitation, is an essential source of information and a major support for all subsequent activities, including these post-mining operations. In many cases, it can be responsibly performed by trained personnel of the mining company, ideally jointly with local or regional NGOs and professionals from state agencies who will also accompany the subsequent phases.

Mining or Exploitation Phase

In principle, similar impacts as described above may occur with the establishment of a full mining operation, although these are a lot more extensive and persistent. In addition, a mining phase could result in a suite of considerably more hazardous and long-lasting impacts. For most environmental compartments, the impact duration is at least as important as the strength of the impact. Mines usually have an operating lifetime of at least 10 years to many decades, a period of direct impact. Such a lengthy span of impact has the potential to leave legacies for centuries or even millennia (see A Concise Review of Mining History).

Atmosphere. Both open-pit and underground mines generate exhausts and considerable amounts of dust, even when properly operated [19]. Dust is generated during aboveground and underground mining, drilling, blasting, and all processes involving transferring, dumping, discharging, crushing, hauling, and processing materials. Depending on local heat, humidity, and wind conditions (local climatology), the impact on the atmosphere may be comparatively large, covering substantial areas with mainly mineral dust and furnace residues (power plants) or even with fine metal aerosols (from smelter operations). Independent of their possible direct toxicity, the settling dusts and aerosols cover plants and soil surfaces, impeding plant respiration and altering the local soil chemistry. Although dusts, aerosols, and other exhausts may travel airborne for up to several thousands of kilometers away from the source, these usually remain within a limited "halo" around the operations. Apart from these, toxic gases may also be released, e.g., sulfur dioxide from mineral sulfides, a major precursor for long-range transport species of key aerosol components (e.g., ammonium sulfate). Very large operations are known to contribute to a great extent to hemispheric pollution, e.g., the Sudbury smelter in Ontario, Canada [21]; the Freiberg smelters in Saxony, Germany [22]; and the Nikel/Zapoljarnyi and Monchegorsk operations on Kola peninsula, Russia [48]. Apart from direct metal emissions, their SO_2 emissions contributed substantially to the atmospheric formation of acidic precipitation and are largely responsible for the related major air pollution with subsequent soil and water pollution in the last decades of the twentieth century [23]. Even the carbon dioxide balance of the operation comes under close scrutiny, since many countries use carbon-trading schemes in order to benefit from implementing smart technology and to penalize big energy wasters. By their very nature, mining activities and equipment can emit high levels of noise and vibration. This usually appears to adversely affect people, including workers, more than most animal species, while no known related impact has been determined on plants.

Hydrosphere. Most mining operations demand comparatively large amounts of energy and water. For some high-volume, high-mass operations, such as coal mining, entire power plants are needed to meet the energy demand of the operations. In addition to cooling water, a water resource is needed in very many parts of the operational stages. The cooling water and its evaporation in cooling ponds or towers may influence the local microclimatology, which involves largely uncritical humidity increases. Open-pit mines may use very large amounts of water for the mining process itself (e.g., high-power water jets, air stripping, machine cooling), and to safeguard infrastructure (e.g., "constant" water spraying to suppress dust generated on haul roads and stockpiles), and particularly for the ore dressing and smelting process (milling, classification and transport as slurries, flotation processes). Additional high water consumption derives from leaching and bioleaching operations. For this reason, water demand itself can pose a major challenge, particularly in dry or semidry environments. The required lowering of groundwater levels around the mining operation (to keep the mine dry, safe, and operable) is another direct imminent impact within the area that is being dewatered throughout the era of active mining. Competition for this water supply with resident people and with terrestrial and aquatic ecosystems can be a contentious issue. In consequence, the water balance at mining sites is altered, and a persistent lowering of water tables or even diminishment of aquifers is often encountered. The described applications lead not only to water losses but potentially to hazardous water contamination. The input of polluted wastewater may directly impact biota – and indirectly, the human body. Acidic mine waters are another big issue, again mostly restricted in their spatial impact – and potentially easy to control. For decades following the 1970s, considerable attention on and inquiry into acid mine drainage (AMD) was triggered by lasting operational and environmental concerns, making it a hot topic [24, 25]. As a consequence of this acidic outflow of water, surface waters (rivers and lakes) and groundwaters may be seriously affected. Sediment pollution needs to be considered [9, 26, 27] since ample examples exist of non-retained mining materials traveling (and contaminating) hundreds of kilometers downstream (e.g., Ok Tedi mine, Papua New Guinea) [28] or of dam failures and subsequent accidents (e.g., Tysa river, Hungary) [29]. Access to clean drinking water remains a challenge in many parts of the world. This issue will remain with us for a long time to come, with a growing world human population exceeding the seven billion humans mark in 2011, and nine billion around the year 2050. While many waterborne pollutants may have a rather limited lifetime (e.g., cyanide from gold mining), persistent organic pollutants (POPs) from drilling operations, ore dressing, and energy conversion and potentially toxic metal species that may reside in aquatic systems for very long periods of time (decades to centuries) pose a lasting challenge.

Such pollutants require particular attention and necessitate safeguarding against any kind of spill, leakage, and loss [30].

Pedosphere. While all mining operations require an initial removal of the natural unconsolidated land surface material (overburden), this is particularly true for open-pit operations. It is common practice to remove and store the nutrient-rich and potentially fertile topsoil separately. The deeper mineral soil material also is removed to free the deposit for active mining and stored separately. This avoids disposal of this subsoil overburden (depending on the operation), which may consist of millions of tons of rock material (usually soft and permeable). In underground mining, the equivalent to this requiring storage is the waste rock from the mining operation. Ore dressing and smelting operations produce partly extensive amounts of tailings, slags, and similar materials that need to be disposed of. Valley filling still appears to be the most sought-after option. As a result and independent of surface or underground mining, comparatively large areas that far exceed the immediate area of the mining facility may become part of the mining operation and of its environmental footprint. Valleys filled or soils covered with such "waste" materials can no longer provide their useful ecological services. Their former habitat function has ended too. While the new morphology and material will attract new life and new ecological equilibria may form (over extended periods of time), the previous ecosystem is no longer functional, and thus, profoundly and permanently impacted. If in addition, the deposited materials contain toxic components, both from the mined raw materials and from chemicals added during the beneficiation processes, these may again further enhance longer-term environmental degradation. Gold-mining legacies with related arsenic toxicity serve as an example [31]. "White mining," the mining of rock salt, further illustrates the challenges: large amounts of impure salt rock debris are being deposited on spoil tips that will persist for centuries. If not covered and not equipped with drain controls (effluent treatment), the easily dissolvable material will deliver excessive amounts of salt into adjacent soils, groundwater, and surface waters. The detrimental effects of excessive amounts of simple mineral salts on plants and many other biota are well known.

Biosphere. With the discussion on ecosystem functions in both the hydrosphere and the pedosphere, it is obvious that the biosphere is strongly impacted too. The first – and often key issue – is habitat loss. This is most certainly the most crucial and critical element of potentially very long-lasting detrimental consequences of mining operations. Although life can re-establish itself even in the most hostile and apparently devastated environments, previous ecosystems may never reestablish. Such consequences could be tolerated if it did not happen at very many places worldwide and if refuge areas did not become increasingly scarcer. Options to protect the biosphere from detrimental impacts are available but often disregarded or considered excessively expensive or demanding. In detail, again a very large array of developments and consequences emerges, depending on biome and local ecosystems. Even if individual species are being extinguished at a specific location, this loss may lead to a domino effect on the web of organisms on all levels – from microbial life via all levels of plants and insects to molluscs, amphibians, fish, reptiles, birds, and mammals. Nutrient supply may become limited due to the mostly fresh rock and overburden materials; the absence of fine materials may further inhibit the growth of higher plants (resettlement). Without further management options, recovery of such sites may take centuries.

Cryosphere. Permanent ice cover and permafrost environments yield potentially attractive mineral resources. These do not only occur at very high latitudes on both hemispheres but also at higher alpine elevations (e.g., Bolivia, Peru). With ongoing global warming, so far mostly inaccessible areas mostly in North America, Siberia, and Greenland as well as in Argentina and Chile become potentially available and feasible for exploration and exploitation. Such environments are extremely sensitive to impacts and will remain sensitive. Their slow biogeochemical cycles retain negative imprints for very long periods of time, and recovery is accordingly extremely slow. Although ice or frozen ground may be compromised, mining will impact exceedingly on the water cycle, the soils (generally very shallow), and the low biodiversity (this low abundance characterizes the rather extreme vulnerability of such environments).

Lithosphere. Even the lithosphere itself can experience a lasting impact, detrimental to future use. Mining subsidence, sinkholes, and drying-up of aquifers are among the most prevalent potential environmental impacts of mining. It is well known that sinkholes may form at the surface over former underground mining operations and that mining subsidence can affect areas of hundreds of square kilometers in size. A notable example (also for a major impact on regional aquifers) is the very densely populated Ruhr area (Ruhrgebiet) in western Germany, where deep coal mining leaves its legacy [32].

Decommissioning and Recultivation Phase

In most modern mines, recultivation commences long before production ceases and the mine is abandoned. An intelligent long-term advance planning may even turn environmental legislation demands into profits. Planning recultivation and handling of environmental issues are key prior to any action. Impressive positive examples can be taken from lignite open-pit mines in the Lusatian basin in Germany (e.g., [33, 34]) and various other places. However, there are still regions where recultivation starts only after decommissioning – if it starts at all. Ever so often, mining companies claim bankruptcy at the end of the operation to save the necessary costs related to recultivation. As leading mining companies, joined in the Global Mining Initiative (GMI; [35]), actively demonstrate their responsibility, a certain fraction of the global mining enterprises still follows a different route – and contributes to the above-mentioned notion of mining being a dirty business. In most countries with a well-developed mining sector, mining companies are forced to put aside funds (usually into trust funds managed by government regulators) that will suffice for recultivation and clean-up of facilities during decommissioning, so that future land use is not compromised. Once the active mining has stopped, all facilities and infrastructures need to be dismantled, removed, and, wherever possible, recycled. Theoretically, the landscape should be returned to its original state prior to the mining-related activities. Water, soils, and biota should be able to recover rapidly. In this phase, however, disturbances are unavoidable, albeit moderate as compared with the active mining phase. The slightly suboptimal reality should be countered by a discussion of some important aspects, namely on the dimensions of scale. An unusual and generally very positive example can be taken from the German superfund site of the Wismut operations, which was a "secret" Russian uranium mining and processing operation in former East Germany. This was one of the world's largest mine closures and remediation projects, "*including five underground mines, and more than 3,700 ha of contaminated areas with ca. 500 million m^3 of solid, radioactively contaminated material*" [36].

Mining enterprises and related activities range from spatially highly restricted small-scale (or artisanal) mining, usually run by local people and often without appropriate training, to very large projects, mostly run by national and international companies with access to highly sophisticated equipment and technology. Such variety cannot be discussed on the same level. Sometimes, small-scale mining may be considerably less environmentally friendly by unit, but if the enterprise remains highly localized, this size restriction at the same time reduces the ecological footprint. At the other end of the scale, a very large operation that manages the site with state-of-the-art techniques may still be making an unsustainably large footprint, simply because of its sheer size. For this reason, it also appears obvious that the boundary conditions of any mine's location play a crucial role in realistically assessing the true impacts of the operation.

The Bottom Line

Mining per se must not be a devastating enterprise as it is ever so often perceived – and undoubtedly often with reason. Only if a comprehensive and open-minded environmental impact assessment is professionally performed from the very beginning and if related recommendations are followed, then the mining operation and its surrounding related activities could be regarded as a rather sustainable enterprise. A skeptic will immediately point out the related assessment costs that may well suppress any entrepreneurial activity and increase financial risks beyond feasibility. That would be a valid argument only if certain boundary conditions are not seen and met.

First and foremost, it has to be acknowledged that the twenty-first century marks the very first human generation that is capable of "seeing the global consequences" of its own activities. Prior to the development and employment of remote-sensing technologies, this awareness was outright impossible. Still, most people only perceive their immediate habitat and often make far-reaching decisions based on that limited worldview. In 2011, the world human population is the largest ever and is predicted to reach nine billion by 2050. This population increase is but one of the many global change challenges: climate change, soil and biodiversity loss, water scarcity, etc., mark a few other hotspots.

Without mining, however, the growing human population would neither be able to improve its standard of living nor maintain its well-being due to the shortage of primary raw materials essential for developing technology and building houses and infrastructure of any kind. One has only to consider the technological demands of modern medicine. The need for new materials emerges only with scientific and technological advancement. Mining will remain a necessity, since even the very best recycling rates cannot provide the amounts needed of various commodities. To avoid or at least drastically curb the damaging side effects of mining and related activities, a different approach deems necessary and paramount – the approach of sustainable mining.

Future Directions – Sustainable Mining

What Is Sustainable Mining?

The strictly regulated mining industry worldwide may strive for but can never attain a completely sustainable mining scenario. Still, when looking at related publications from the mining industry and authorities (e.g., [20, 37–40]), the notion of sustainable mining has taken a stronghold and increasingly focuses on the social and environmental issues. Sustainability has been clearly defined as having the social, the economic, and the environmental perspective in view [41]. Yet, there is a need for a strong practical bias towards environmental issues when thinking about sustainable mining. Without a "healthy" environment, there would be a rather grim future for both social and economic issues. Rajaram et al. [42] provide a helpful discussion in this respect while shying away from a distinct definition.

Simply spoken, sustainable mining is the kind of mining activity that does not compromise the future long-term well-being of people on or near sites of earlier mining. There may be a discussion on what constitutes "well-being." Hence, a pragmatic, less philosophical approach is suggested by defining "well-being" as the state of a human being where basic social and health needs are met. Since these demand a healthy environment, this three-part perspective is essential to sustainability. How difficult this may be in detail, however, has been addressed by Marker et al. [43] with various multi-scale examples, particularly from the developing world. They argue for a concept of an "*ideal sustainability model as one that minimizes negative environmental impact and maximizes benefits to society, the economy and regional/national development.*" They also acknowledge the long-term character of such an approach, if taken seriously and if broad acceptance is to be achieved.

As a result, the near future will most likely see both conventional mining and also emerging new methods and technologies. These may include phyto-mining and the use of microbial assemblies to access and bring forth desired commodities without large rock and material movement. It will include in situ leaching and in situ processing of ores and will avoid the buildup of waste-rock piles and tailings deposits. It will also see a changing approach from the focus on a single or limited commodity to a broader and more long-term view that avoids producing "wastes" and rather safeguards and leaves future options open. At the same time, however, more surface operations that exploit increasingly lower concentrations of the commodities – with all potential risks involved – are seen. Both underground and surface (open-pit) mining can be done without compromising the environment for future generations. To understand such a claim, an even more complex vision needs to be developed.

The Complete Budget

The term "waste" is purposely accentuated here in quotation marks in order that those unneeded materials not be regarded as waste in the literal sense but rather as a potential future resource. It may be equally necessary not only to look at the entire mining business as an enterprise that will deliver commodities but also to address the issue in a much broader context.

Just like a water reservoir can be seen as a constructed body for the provision of drinking water or water for industrial purposes alone, it can also be seen as a multifunctional construct that potentially provides hydro energy, flood protection, fish-farming, recreation, and more opportunities and services. Obviously, these additional services may deliver a significant benefit to society. Can mining be seen and interpreted in a similar fashion? It can, although such a perspective demands a rather radical redefinition of the role of mining.

The paradigm change needed demands a complex and holistic long-term view, where a mining company plays a role as a service provider for society at large and not just as an independent private business. As a consequence, a much closer and partnership-based relation would be developed between all stakeholders: the company, the government (local, regional, or national), and the regional populace. A strategy developed by the chemical industry that serves as an example (however, which would need to be adapted to the mining sector) is the concept of eco-efficiency analysis [44, 45]. Adapted for use by the mining industry, a company would benefit from delivering additional services up to the decommissioning and possibly the rehabilitation phases. It would earn its money not only through selling a commodity to the global market but also through the complex added values, hence improving the socioeconomic situation in the region (which is often done already). It might develop post-enterprise industrial activities to ensure the subsequent benefit for the region, and could plan and establish the rehabilitation activities, based not on the minimum but the maximum possible requirements. This includes looking at mining wastes as a potential future commodity that needs to be safeguarded for easy, energy-efficient, and safe retrieval at a later time. All of these added activities generate additional costs, although if done properly, these may save a lot of future costs that are paid by the tax payer and easily excel the monetary benefit of the mining operation itself. One visionary example further illustrates this point where a back-end approach is taken. It is the complex knowledge of an ore body or reserve and its setting that drives the planning for exploiting the mine. The planning is not driven by momentary market prices (that contradict maximum resource efficiency) but by the objective and longer-term necessities and requirements for an efficient, safe, and complete utilization of all commodities in that deposit. "Waste" could be used as construction and building materials, and all toxic components could be extracted as by-products, recycled, or stored in a safe manner to serve future generations as a secondary resource.

Thus, the aim is to establish a long-term partnership and win-win situation for the benefit of all – the company, the employers and residents, and the environment – and to further the development of the region. It basically turns from a single business economy approach to a long-term perspective of political economics. With the most likely future political developments (long-term perspective) in mind, this will translate to international political economy rather than national economy.

The downside, at least as it may be perceived by a company, clearly means a much longer planning phase, the demand for early and truly open communication with all stakeholders (including risk communication), and the necessity of a much more transparent operation throughout as compared to the prevalent current standards. There are quite a few "walls to surmount" and even more prejudice and traditional concepts to overcome. Particularly, the mining industry is still largely characterized by a rather conservative approach.

The benefits are obvious: mining companies and related enterprises can no longer be perceived as obscure omnipotent malevolent entities, interested in basically nothing but the provision of industry with commodities, but as badly needed and responsible partners. Sustainable mining operations will be involved not only in the necessary acquisition and refinement of raw materials but also in the recuperation and delivery of the exploited area to future generations without compromising that future.

Such a vision is nothing short of revolutionary. Yet, it may need truly revolutionary attempts to successfully face the global challenges and to support a still growing human population – without waging wars and without turning a blind eye to extreme socioeconomic disparity.

Bibliography

Primary Literature

1. Schneider J (1998) Environmental impact of marine mining. N Jahrb Geol Paläont Abh 208:397–412
2. Chamley H (2003) Geosciences, environment and man. In: Chamley H (ed) Developments in earth and environmental sciences, 1. Elsevier, Amsterdam, 527 p

3. Ellis D (1989) Environments at risk. Case histories of impact assessment. Springer, Berlin/New York, 329 p
4. IRMA (2011) Documents. The initiative for responsible mining assurance. http://www.responsiblemining.net/documents.html. Accessed 8 Sept 2011
5. Kausch P, Ruhrmann G (2001) Environmental management. Environmental impact assessment of mining operations. Logabok, Köln, 133 p
6. Bednarik RG (1992) Early subterranean chert mining. Artefact 15:11–24
7. Dart RA (1967) The antiquity of mining in Southern Africa. S Afr J Sci 63(6):264–267
8. Dart RA, Beaumont PB (1968) Ratification and retrocession of earlier Swaziland iron ore mining radiocarbon datings. S Afr J Sci 64(6):241–246
9. Matschullat J, Ellminger F, Agdemir N, Cramer S, Liessmann W, Niehoff N (1997) Overbank sediment profiles – evidence of early mining and smelting activities in the Harz mountains, Germany. Appl Geochem 12:105–114
10. Sparks DL (2005) Toxic metals in the environmental: the role of surfaces. Elements 1(4):193–197
11. Klappauf L, Linke FA, Brockner W, Heimbruch W, Koerfer S (1990) Early mining and smelting in the Harz region. In: Pernicka E, Wagner GA (eds) Archaeometry, vol 90. Birkhäuser Verlag, Basel, pp 77–86
12. Rebrik BM (1987) Geologie und Bergbau in der Antike. Deutscher Verlag für Grundstoffindustrie, Leipzig, 183 p
13. Rosman KJR, Chisholm W, Hong S, Candelone JP, Boutron CF (1997) Lead from Carthagian and Roman Spanish mines isotopically identified in Greenland ice dated from 600 B.C. to 300 A.D. Environ Sci Technol 31:3413–3416
14. MHN (1997) The mining history network, http://projects.exeter.ac.uk/mhn/. Accessed 8 Sept 2011
15. Diamond J (2005) Collapse. How societies choose to fail or survive. Penguin, London, 575 p
16. Down CG, Stocks J (1977) Environmental impact of mining. Applied Science, London, 380 p
17. Carlowitz HC von (1713) Sylvicultura oeconomica. Anweisung zur wilden Baum-Zucht. Reprint of the 1713 ed Leipzig, Braun, revised by Klaus Irmer and Angela Kießling, TU Bergakademie Freiberg and Akademische Buchhandlung, Freiberg 2000, ISBN 3-86012-115-4; Reprint of the 2nd ed from 1732, Verlag Kessel, ISBN: 978-3-941300-19-4
18. Grober U (2010) Die Entdeckung der Nachhaltigkeit. Kulturgeschichte eines Begriffs. Kunstmann Antje GmbH, 300 p
19. Plumlee GS, Ziegler TL (2005) The medical geochemistry of dusts, soils and other Earth materials. In: Sherwood Lollar B (ed) Environmental geochemistry. In: Holland HD, Turekian KK (ser eds) Treatise on geochemistry, vol 9, issue 7, pp 263–310
20. PDAC (2009) e3plus – a framework for responsible exploration, 34 p. http://www.pdac.ca/e3plus/. Accessed 8 Sept 2011
21. Gunn JM (ed) (1995) Restoration and recovery of an industrial region, Environmental management. Springer, New York, 358 p
22. Ilgen G, Fiedler HJ (1990) Smelter smoke damage at Freiberg in the 19th century, and its study by Professors Reich (Freiberg) and Stöckhardt (Tharandt) II Explaining the causes of damage by agricultural chemistry methods. Wiss Z TU Dresden 29(6):115–118
23. Last FT, Watling R (1991) Acidic deposition – its nature and impacts. Proc Royal Soc Edinburgh Section B (Biological Sciences) 97:343
24. Blowes DW, Ptacek CJ, Jambor JL, Weisener CG (2005) The geochemistry of acid mine drainage. In: Sherwood Lollar B (ed) Environmental geochemistry. In: Holland HD, Turekian KK (ser eds) Treatise on geochemistry, vol 9, issue 5, pp 149–204
25. Singer PC, Stumm W (1970) Acidic mine drainage: the rate-determining step. Science 167(3921):1121–1123
26. Knittel U, Klemm W, Greif A (2005) Heavy metal pollution downstream of old mining camps as a result of flood events: an example from the Mulde river system, eastern part of Germany. Terr Atmos Ocean Sci 16(4):919–931
27. Ridgway J, Flight DMA, Martiny B, Gomez-Caballero A, Macias-Romo C, Greally K (1995) Overbank sediments from central Mexico: an evaluation of their use in regional geochemical mapping and in studies of contamination from modern and historical mining. Appl Geochem 10:97–109
28. Pernetta JC (1988) Potential impacts of mining on the Fly river, UNEP 99, 191 p
29. Hum L, Matschullat J (2003) Gold kann schmutzig sein. Welche längerfristigen Auswirkungen hatte das Unglück bei Baia Mare auf die Theiss? In: Unland G, Herzog P (eds) Der Bergbaubezirk Baia Mare, Rumänien. Eine komplexe Betrachtung der Lagerstätte, des Bergbaus, der Aufbereitung sowie der Umweltfolgen. TU Bergakademie Freiberg, Freiberg
30. Goudie A (2006) The human impact on the natural environment, 6th edn. Blackwell, Oxford, 357 p
31. Deschamps E, Matschullat J (2011) Arsenic: natural and anthropogenic. In Bundschuh J, Bhattacharya P (ser eds) Arsenic in the environment, vol 4. CRC Press, Balkema, 209 p
32. Bell FG, Stacey TR, Genske DD (2000) Mining subsidence and its effects on the environment: some differing examples. Environ Geol 40(1–2):135–152
33. Hüttl RF (1998) Ecology of post-mining landscapes in the Lusatian lignite mining district, Germany. In: Fox HR, Morre HM, McIntosh AD (eds) 4th International conference of the internat affiliation of land reclamationists. Balkema, Nottingham, UK
34. Krümmelbein J, Horn R, Raab T, Bens O, Hüttl RF (2010) Soil physical parameters of a recently established agricultural recultivation site after brown coal mining in East Germany. Soil Tillage Res 111(1):19–25
35. Littlewood G (2000) The global mining initiative. Address to Mining 2000, Melbourne September 20. www.icmm.com/document/104. Accessed 8 Sept 2011
36. Paul M, Mann S (2010) Environmental clean-up of the East German uranium mining legacy: discussion of some key experiences made under the Wismut remediation program. In: Lam E, Rowson J, Özberk E (eds) Uranium 2010 – Proc 3 rd Internatational conference uranium, vol II, 15–18 Aug, Saskatoon, Saskatchewan, Canada, pp 481–493

37. AusIMM (2011) Australasian institute of mining and metallurgy. http://www.ausimm.com.au/. Accessed 8 Sept 2011
38. CSIRO (2011) Sustainability. Commonwealth Scientific and Industrial Research Organisation. http://www.csiro.au/science/Sustainability.html. Accessed 8 Sept 2011
39. Mining Association of Canada (2011) Towards sustainable mining. http://www.mining.ca/www/Towards_Sustaining_Mining/index.php. Accessed 8 Sept 2011
40. PDAC (2007) Prospectors and developers association of Canada http://www.pdac.com.br/2007/english/index.htm. Accessed 8 Sept 2011
41. United Nation (1987) Report of the World commission on environment and development: our common future. http://www.un-documents.net/wced-ocf.htm (Brundtland Commission)
42. Rajaram V, Dutta S, Parameswaran K (2005) Sustainable mining practices: a global perspective. CRC Press, Baco Raton, 370 p
43. Marker BR, Petterson MG, McEvoy F, Stephenson MH (eds) (2005) Sustainable minerals operation in the developing world. Geological Society Special Publication 250, 249 p
44. Saling P, Kicherer A, Dittrich-Krämer B, Wittlinger R, Zombik W, Schmidt I, Schrott W, Schmidt S (2002) Eco-efficiency analysis by BASF: the method. Int J Life Cycle Assess 7(4):203–218
45. Shonnard DR, Kicherer A, Saling P (2003) Industrial applications using BASF eco-efficiency analysis: perspectives on green engineering principles. Environ Sci Technol 37(23):5340–5348
46. Agricola G (1556) De re metallica. Libri XII. English language version from 1912 by Hoover H and Hoover LH; ISBN 0-486-60006-8; 650 p
47. IIED (2002) Breaking new ground: mining, minerals and sustainable development. 462 p. http://www.iied.org/sustainable-markets/key-issues/business-and-sustainable-development/mmsd-final-report. Accessed 8 Sept 2011
48. Reimann C, Äyräs M, Chekushin V, Bogatyrev I, Boyd R, Caritat P de, Dutter R, Finne TE, Halleraker JH, Jæger Ø, Kashulina G, Lehto O, Niskavaara H, Pavlov V, Räisänen ML, Strand T, Volden T (1998) Environmental geochemical atlas of the Central Barents Region. NGU-GTK-CKE Special Publication, Geological Survey Norway, Trondheim, Norway, 745 pp. http://www.schweizerbart.de/publications/detail/isbn/9783510652631. Accessed 8 Sept 2011
49. Adriano DC (2001) Trace elements in terrestrial environments. Biogeochemistry, bioavailability and risks of metals. 2nd ed. Springer, New York, 867 p

Books and Reviews

Abdelouas A (2006) Uranium mill tailings: geochemistry, mineralogy, and environmental impact. Elements 2(6):335–341

Breitkreuz C, Drebenstedt C (eds) (2009) Sustainable mining and environment – a German – Latin American perspective. TU Bergakademie Freiberg, Freiberg

Einaudi MT (2000) Mineral resources: assets and liabilities. In: Ernst WG (ed) Earth systems: processes and issues, vol 23. Cambridge University Press, Cambridge, pp 346–372

Figueiredo BR (2000) Minérios e ambiente. Editora da Unicamp, Coleção Livro-Texto, 401 p

Fubini B, Fenoglio I (2007) Toxic potential of mineral dusts. Elements 3(6):407–414

Hüttl RF, Heinkele T, Wisniewski J (1996) Minesite recultivation. Springer, New York, 172 p

Maskall J, Whitehead K, Thornton I (1995) Heavy metal migration in soils and rocks at historical mining sites. Environ Geochem Health 17:127–138

Mining, People and the Environment (online magazine) http://www.mpe-magazine.com/. Accessed 8 Sept 2011

Morin G, Calas G (2006) Arsenic in soils, mine tailings, and former industrial sites. Elements 2(2):97–102

Sharma AK (no year) Scientific and sustainable mining. www.fedmin.com/html/goapaper.pdf. Accessed 8 Sept 2011

Woodward J, Place C, Arbeit K (2000) Energy resources and the environment. In: Ernst WG (ed) Earth systems: processes and issues, vol 24. Cambridge University Press, Cambridge, pp 373–401

Mining Industries and Their Sustainable Management

Sandip Chattopadhyay[1], Devamita Chattopadhyay[2]
[1]Tetra Tech Inc, Cincinnati, OH, USA
[2]CH2M Hill, Dayton, OH, USA

Article Outline

Glossary

Abandoned mines Mines for which the owner cannot be found, or for which the owner is financially unable or unwilling to carry out cleanup. They may pose environmental, health, safety, and economic problems to communities, the mining industry, and governments in many countries.

Acid (rock or mine) drainage Many metal ore bodies and coal deposits contain significant quantities of sulfide minerals – often including the ore minerals themselves. When such minerals are brought to the surface, they react chemically with air and water producing sulfuric acid, which may dissolve other minerals containing potentially toxic elements. This *acid drainage* from coal and metal mining around the world can pollute water and the surrounding land, affecting plant and animal life. Acid drainage is known as *acid mine drainage* when it is closely associated with mining activities, and *acid rock drainage* when this phenomenon occurs naturally, without human intervention. Both phrases are in common use, although particular stakeholder groups have particular preferences related to the controversial nature of this issue.

Acidophile An organism that thrives in a relatively acidic environment.

Ammonification The biochemical process whereby ammoniacal nitrogen is released from nitrogen-containing organic compounds.

Amorphous Irregular, having no discernible order or shape. Rocks or minerals that possess no definite crystal structure or form, such as amorphous carbon.

Bioleaching Extraction of metal from solid minerals into a solution is facilitated by the metabolism of certain microorganisms.

Biomining Extraction of specific metals from their ores through biological means, usually bacteria. It is an actual economical alternative for treating specific mineral ores, involving percolation and agitation techniques.

Community The people living around the mine who are directly affected (both positively and negatively) by the mine's activities.

Contaminated land/water Land/water containing concentrations of potentially toxic materials (organic or inorganic) elevated above the natural background concentrations in a particular area. In relation to mining, land or water contamination may occur through fuel spills, run-off from waste rock dumps, leaks from tailings impoundments, windblown dust from tailings and waste rock, smelter emissions, and drainage from mine workings. Contaminated groundwater is caused by the seepage of contaminated waters into aquifers.

Crystalline A substance in which the constituent atoms, molecules, or ions are packed in a regularly ordered, repeating three-dimensional pattern.

Denitrification A microbially facilitated process of nitrate reduction that may ultimately produce molecular nitrogen through a series of intermediate gaseous nitrogen oxide products.

Dissimilatory reduction Sulfate-reducing bacteria reduce sulfate in large amounts to obtain energy and expel the resulting sulfides as waste; this is known as dissimilatory sulfate reduction. They are anaerobes, which use sulfate as the terminal electron acceptor.

Electrowinning The recovery of metal by electrolysis. An electric current is passed through a solution containing dissolved metals, and this causes the metals to deposit on a cathode.

Extractant An immiscible liquid used to extract a substance from another liquid.

Gypsum A sedimentary rock consisting of hydrated calcium sulfate.

Heap leaching To dissolve minerals or metals out of an ore heap using chemicals. For example, a cyanide solution percolates through crushed ore heaped on an impervious pad or base pads during heap leaching of gold.

Macrophyte An aquatic plant that grows in or near water and is either emergent, submergent, or floating. They provide cover for fish and substrate for aquatic invertebrates, produce oxygen, and act as food for some fish and wildlife.

Mesophile An organism that grows best in moderate temperature (typically between 15°C and 40°C). The term is mainly applied to microorganisms. It is also used to describe mesophilic anaerobic digestion, which takes place optimally around 37–41°C or at ambient temperatures between 20°C and 45°C, where mesophiles are the primary microorganisms present.

Methanogenesis Production of methane by biological processes that are carried out by methanogens. A methanogen is a single-celled microorganism and is a member of the Archaea. Archaea are unique because unlike most life on Earth that rely on oxygen and complex organic compounds for energy, Archaea rely on simple organic compounds (e.g., acetate) and hydrogen for energy.

Mining life cycle The processes of exploration, mining development, extraction, processing, refining, smelting, and mine closure.

Ore Mineral-bearing rock that can be mined and treated profitably under the existing economic conditions, or those conditions which are deemed to be reasonable.

Reclamation Process of converting derelict land (land that requires intervention before beneficial use) to usable land and may include engineering as well as ecological solutions.

Remediation Environmental cleanup of land and water contaminated by organic, inorganic, or biological substances.

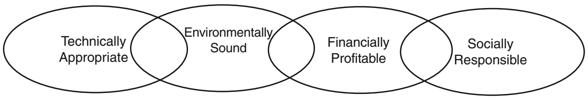
Restoration Seeks to artificially accelerate the processes of natural succession by putting back the original ecosystem's function and form.

Tailings Mineral wastes produced from the processing operations after the valuable minerals have been extracted.

Waste rock The mineral wastes produced during mine development – including overburden and barren rock – and those parts of an ore deposit below the economic cutoff grade. Often, and particularly in some metal deposits, the waste rock may contain sufficient sulfide mineral concentrations to generate long-term acid drainage problems.

Definition of the Subject and Its Importance

Mining, minerals, and metals are important to the economic and social development as they are essentials for modern living. However, supplies of minerals, such as coal, are limited, and sustainable management of natural resources requires the maintenance, rational and enhanced use as well as a balanced consideration of ecology, economy, and social justice. Mining industry's recognition and acceptance of its sustainable development is growing. In the mining and metals sector, this means that investments should be:



Leaders in the global mining business community have begun to assign a new strategic significance to the term sustainability. No longer confined to the economic realm, sustainability embraces a broad spectrum of organization characteristics related to social and environmental responsibility. Profitability alone is inadequate as a measure of success, and that many of the nonfinancial concerns associated with sustainability are fundamental drivers of long-term shareholder value. Failure to recognize these strategic issues threatens the very survival of a business enterprise.

Large amounts of material are involved in large-scale mining and minerals extraction. The problems arising from the change in the chemistry of million tons of natural ore during the processing steps (such as, grinding, calcining, and roasting operations) and their resultant bioavailabilities are not well understood. Mining produces large volumes of waste, and decisions regarding waste handling and other aspects of operations are often difficult and expensive to reverse; they need to be made correctly initially through mine closure planning. Another challenge is the environmental legacy left by mining. The environmental issues of current mining operations are daunting enough. But in many ways far more troubling are some of the continuing effects of past mining and smelting. The loss of biodiversity is the other great challenge of mining sustainability. The loss of biodiversity is an irreversible loss. Conservation practices that guarantee a minimum impact on biodiversity must be adopted and implemented.

As important as the methods of mining and beneficiation is how the minerals are used for sustainable development. An integrated approach for the use of minerals must be developed. Current patterns of minerals used raise concerns about efficiency and the need for more equitable access to resources worldwide. The environmental and health impacts of different mineral products in use need to be carefully managed.

Where the risks associated with use are deemed unacceptable or are not known, the costs associated with using certain minerals may outweigh the benefits.

The importance of sustainable development in mining industry includes actions at all levels to:

- Support efforts to address the environmental, economic, health, and social impacts of mining throughout the life cycle, including workers' health and safety. A range of partnerships, furthering existing activities at the national and international levels, among interested governments, intergovernmental organizations, mining companies and workers, and other stakeholders to promote transparency and accountability for sustainable mining and minerals development.
- Enhance the participation of stakeholders, including local and indigenous communities, to play an active role in minerals, metals, and mining development throughout the life cycles of mining operations. This includes efforts after mine closure for rehabilitation purposes, in accordance with national regulations and taking into account significant transboundary impacts.
- Foster sustainable mining practices through the provision of financial, technical, and capacity-building support to developing countries and countries with economies in transition for the mining and processing of minerals. This includes small-scale mining, and, where possible and appropriate, improve value-added processing, upgrade scientific and technological information, and reclaim and rehabilitate degraded sites.

Introduction

Mining takes a significant toll on the environment. The intensity of the environmental impact depends on what is being mined, where and how. Unless it is meticulously planned and carefully executed, mining can devastate lands, pollute and deplete water resources, denude forests, wipe out wildlife, and defile the air.

The history of mining has been one of boom and bust periods that were a balance between available natural resources, production costs, and the ability to sustain profitability. In recent years, the mining industry leadership has emphasized that the modern-day mining business has become more complex with the advent of stricter regulations, better-informed stakeholders, and closer in-depth scrutiny of current and/or proposed operations. Increased emphasis is being placed on the importance of sustainable operations that are focused on effectively integrating environment, social, and economic impacts.

The quantity of mining waste produced fluctuates yearly, and as the individual mines and quarries manage their wastes according to local conditions, there are no definitive statistics. After being removed, *waste rock*, which often contains acid-generating sulfides, heavy metals, and other contaminants, is usually stored above ground in large free-draining piles. This waste rock and the exposed bedrock walls from which it is excavated are the primary sources of pollution caused due to mining. The US mining industry produces approximately 8,000,000 tons per year (t/year) of process residue that may contain hazardous species as well as valuable by-products. These process residues are generated by smelter off-gas cleaning at approximately 5,500,000 t/year, and baghouse dust and wastewater treatment at approximately 2,100,000 t/year [24]. Comparable statistics was obtained for other countries. It was estimated that 96.4 million tons of mining and quarrying waste was produced in 2004 in the UK [4]. The Canadian mineral industry typically generates one million tons of waste rock and 950,000 t of *tailings* per day, totaling 650 million tons of waste per year [29]. The right technology may be able to recover marketable by-products from process residues to generate revenue and reduce disposal costs for the mining industry. The process residue that cannot be reused can impact the water resources, and the effect may be manifest throughout the life cycle of the mine and even long after mine closure.

Mining Life Cycle

Minerals are nonrenewable resources, and the mines have finite lives. Mining represents a temporary use of the land, and during this temporary use of the land, the *mining life cycle* can be divided into the following stages: exploration, development, extraction and processing, and mine closure.

Exploration is the work involved in determining the location, size, shape, position, and value of an ore body using prospecting methods, geologic mapping, and field investigations, remote sensing (aerial and satellite-borne sensor systems that detect ore-bearing rocks), drilling, and other methods. Building access roads to a drilling site is one example of an exploration activity that can cause environmental damage.

The development of a mine consists of several principal activities: conducting a feasibility study, including a financial analysis to decide whether to abandon or develop the property; designing the mine; acquiring mining rights; filing an Environmental Impact Statement (EIS); and preparing the site for production. An example of site preparation impacting the environment is removal of the overburden (surface material above the ore deposit that is devoid of ore minerals) by excavation prior to mining.

Extraction is the removal of ore from the ground on a large scale by one or more of three principal methods: surface mining, underground mining, and in situ mining (extraction of ore from a deposit using chemical solutions). After the ore is removed from the ground, it is crushed so that the valuable mineral in the ore can be separated from the waste material and concentrated by flotation (a process that separates finely ground minerals from one another by causing some to float in a froth and others to sink), gravity, magnetism, or other methods, usually at the mine site, to prepare it for further stages of processing. The production of large amounts of waste material (often very acidic) and particulate emission have led to major environmental and health concerns with ore extraction and concentration. Additional processing separates the desired metal from the mineral concentrate.

The closure of a mine refers to the cessation of mining at a site. Planning for closure is often required to be ongoing throughout the life cycle of the mine and not left to be addressed at the end of operations. It involves completing a *reclamation* plan and ensuring the safety of areas affected by the operation, for instance, sealing the entrance to an abandoned mine. The Surface Mining and Control Act of 1977 states that reclamation must “restore the land affected to a condition capable of supporting the uses, which it was capable of supporting prior to any mining, or higher or better uses.” *Abandoned mines* can cause a variety of health-related hazards and threats to the environment, such as the accumulation of hazardous and explosive gases when air no longer circulates in deserted mines, use of these mines for illegal residential or industrial dumping, and others. Many closed or abandoned mines have been identified by federal and state governments and are being reclaimed by both industry and government.

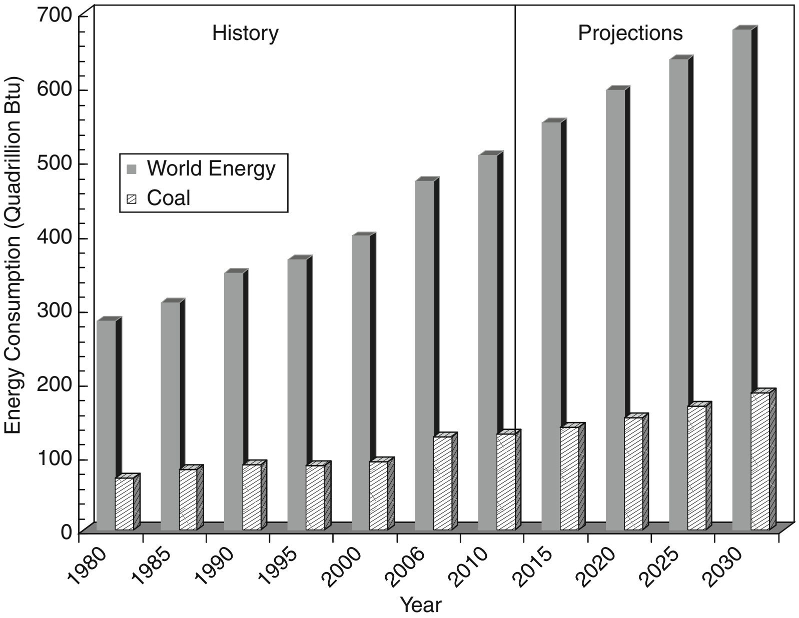
Coal Mining

Mining can affect both air and water. A fossil fuel that has been mined intensively is coal. As an example of environmental impact, effects of coal mining are presented below.

Coal is and will continue to be a crucial element in a modern, balanced energy portfolio, providing a bridge to the future as an important low cost and secure energy solution to sustainability challenges. A review of the Annual Energy Outlook 2010 report from the Energy Information Administration (EIA) indicates that world coal consumption will increase by 56%, from 132 quadrillion BTU in 2007 to 206 quadrillion BTU in 2035. Figure 1 presents the data for coal consumption [6] versus total energy consumption [17] between 1980 and 2030.

The growth rate for coal consumption is uneven, averaging 1.1% per year between 2007 and 2020 and 2.0% per year between 2020 and 2035. The slower growth rate for the earlier period is largely from a decline in coal consumption in 2009 during the global economic recession. With economic recovery, world coal consumption rebounded, and it is expected to return to its 2008 level by 2013. The US energy demand is expected to grow at an average annual rate of 1.1% for the next 25 years. Among the largest consumers of energy are China and India, and together they accounted for about 10% of the world’s total energy consumption in 1990 and 20% in 2007. In a separate report by Waddell and Pruitt [31], it is projected that coal utilization will increase worldwide by 44% by 2025.

The process of coal excavation consists of several cycles: the cycle of excavation, disposal, and recultivation. The first step is reconnaissance of the terrain (preliminary survey of ground) that has been identified for mining. Preparation of the terrain may necessitate demolition of constructions, felling of trees, relocation of watercourses, roads, etc. This is usually followed by excavation of the overburden and then transporting and disposing the material using specialized equipment (mining machinery). In present days, bucket wheel excavators, conveyor belts, and



Mining Industries and Their Sustainable Management. Figure 1
Global energy consumption 1980–2030

stackers are mainly used. When the coal seam appears, such a seam is excavated and transported to the crushing plant.

Some of the important characteristics of coal excavation are listed below:

- Notwithstanding the occupancy of large areas and the change in the purpose of the land, end of mining will result in bringing back large portions of the used areas to the previous state. It means that though certain areas are temporarily occupied by mining activities, for example, for a period of 10 or more years, they can be reused as before following this period.
- Area occupied by the external dump, although it is generally higher than the terrain before mining activities, can be brought back to the state so that it can be reused.
- At the end of mining activities, the mine becomes the owner of large areas of agricultural land and forests. The owner of an agricultural land and forest is obliged to work the soil and control forest husbandry.

Pollution from Coal Mining

The process of coal excavation causes deterioration of the original morphological and pedological structure of the terrain and soil, and results in the release of harmful substances and/or mineral dust into the air. Such a release is primarily related to the deterioration of the upper seam structure during mining operations. The negative impact may occur as a result of the excavation of the upper seam and its inadequate disposal, and of mixing the upper seam with the lower one, as well as other barren materials. The impact of lignite excavation also represents a potential contamination of the upper seam due to the precipitation of dust from the air.

In addition to the aforementioned impacts, mining in a particular area will result in an increase in the area population not only due to the increased number of mine workers but also due to the development of related industries. Such development will often result in disappearance of the arable upper seam due to the building up of infrastructure facilities (roads, railroad tracks, waterways, industrial areas, etc.) and of the change in the purpose of the soil in the vicinity of the mine.

Mining can cause physical disturbances to the landscape, creating eyesores such as waste-rock piles and open pits. Such disturbances may contribute to the decline of wildlife and plant species in an area. In addition, it is possible that many of the pre-mining surface features cannot be replaced after mining ceases.

Mine subsidence (ground movements of the earth's surface due to the collapse of overlying strata into voids created by underground mining) can cause damage to buildings and roads.

Opencast coal mining leads to the disturbance of geological layers where different sedimentary products are being mixed, which means that a completely new, anthropogenic land, i.e., substratum, is created, without any resemblance to the original land, which is called a deposol. This is an example of visual pollution where the characteristics of the landscapes' appearance have changed due to mining. The opencast coal exploitation results in morphological modifications of the terrain, such as the creation of large-scale depressions and the formation of outside overburden dumps.

Pollution of Air

Coal mines emit constituents that cause or contribute significantly to air pollution and that may reasonably be anticipated to endanger public health and welfare. In each stage of mining, from exploration to ore recovery to downstream processing, there is the potential for air quality impacts due to emissions of particulates (dust, diesel, and silica). The monochromatic appearance of the mine areas is due to generation of large quantities of fugitive dusts during mining operations. Coal mining areas are black, bauxite and iron-ore rich regions are red, while limestone gives a chalky white hue. Dust results from blasting, handling, processing, or transporting of soil and rock or can arise from bare or poorly vegetated areas in combination with air movements. It is one of the most visible, invasive, and potentially irritating impacts of mining, and its visibility often raises concerns. Many dusts contain metals which are potentially hazardous, and are known to cause certain diseases. It has the potential to severely affect flora and fauna near the mine and to impact the health of mine workers and local residents. Dust also affects the agricultural productivity of the area. The level of dust generated, its behavior (particle size, density, travel distance), and types of health and environmental risks depend on many factors including mine type, local climate, topography, working methods, types of equipment used, the mineralogy and metallurgical characteristics of some ores, and the land use of the area around the mine.

- EIA reported that estimated recoverable reserves of coal in USA stand at 275 billion tons, an amount that is greater than any other nation in the world. All of the energy growth forecasts have major carbon dioxide (CO_2) emission consequences. The world energy outlook [32] indicated that CO_2 emissions will increase from 26.6 gigatonnes (Gt) in 2005 to 34.1 Gt in 2015 and 41.9 Gt in 2030 [10]. In the Alternative Policy Scenario [32], CO_2 emissions rise to 31.9 Gt in 2015 and to 33.9 Gt in 2030. EIA of the US Department of Energy (DOE) in its International Energy Outlook (IEO) 2007 forecasted that CO_2 emissions in the Reference Case will increase from 26.9 Gt in 2004 to 33.9 Gt in 2015 and to 42.9 Gt in 2030. In the past, global CO_2 emissions rose from 23.5 Gt in 2000 to 27.1 Gt in 2005 [11], which is an increase of over 15%. Based on this, one can say that the forecasts made by IEA and IEO are optimistic.
- Methane is the second most emitted greenhouse gas after CO_2, and is more than 20 times more potent than CO_2 in terms of its heat-trapping capabilities. Methane is a by-product of coalification, the process by which organic materials convert into coal. It is stored throughout the surrounding rock strata in varying sized pockets and, due to the greater overburden pressures, often increases in concentration the deeper the coal seam. Because methane can create hazardous working conditions for miners, it must be removed from underground mines. While methane escapes during the processing, transport, and storage of coal, 90% of the emissions come from the actual coal mining process from all three categories of coal mines: surface mines, underground mines, and abandoned mines.

In 2004, methane accounted for 14.3% of the total anthropogenic greenhouse gas emission load. Since pre-industrial times, methane has contributed to 22.9% of the greenhouse gas load in the Earth's atmosphere. Methane is more abundant in the Earth's atmosphere now than it has been at any time during the past 400,000 years, and the average atmospheric concentration of methane has increased 150% since 1750 due to human activities [25]. The US EPA [27] has concluded that recovering methane from coal mines would significantly reduce the amount of greenhouse gases emitted into the atmosphere because every ton of methane recovered is equal to approximately more than 20 t of carbon dioxide emissions. While methane is a danger in coal mines, it can, if captured, be a valuable commodity: natural gas. To date, underground mines have the most potential to generate a profitable amount of methane. Mines that have adopted a "methane to markets" program as promoted by the US EPA have been successful at methane capture and sale because research has been developing for over three decades to create the best technologies.

- Particulate matters (PM), including total suspended particulates, PM_{10} (particle size less than 10 μm), and $PM_{2.5}$ (particle size less than 2.5 μm), are released during coal mining operations. Coal mining activities that can lead to the release of particulate matter are blasting, truck loading, bulldozing, dragline operation, vehicle traffic, grading, and storage piles.
- Volatile organic compounds (VOCs), including nonmethane organic compounds, are often vented along with methane, from coal mining operations. VOCs are precursors to ground-level ozone, a criteria air pollutant, and are regulated under the Clean Air Act.
- Nitrogen oxides (NOx) are a group of gases that are known to be ground-level ozone and $PM_{2.5}$ precursors and that include nitrogen dioxide (NO_2), a criteria pollutant. Sources of NOx at coal mines include fugitive emissions from overburden and coal-blasting events, tailpipe emissions from mining equipment, point source emissions from stationary engines, coal-fired hot water generators, and natural gas–fired heaters.
- For some pollutants, such as sulfur dioxide, the minerals processing industry is the largest source of emissions.
- Certain kinds of ore, for example uranium, are radioactive, and pose health problems to workers and adjacent communities. Radioactivity follows with whatever radioactive materials escape into the atmosphere or water, and accompanies this material to its final destination. Besides targeted minerals that are radioactive, the host rock may be radioactive, and pose problems for workers. Radioactivity is also present in a gas, radon, which can cause problems for workers in underground mines, if it is present.

Impacts of Dust

The impacts of dust, especially from coal mine area are included below:

Effects on animals: Coal dust is a tumorigenic agent in experimental animals. Coal dusts are associated with lymphomas and, at the higher dose, adrenal cortex tumors in rats exposed to 6.6–14.9 mg/m^3 for 6 h/day intermittently for 86 weeks.

Effects on humans: Coal dust causes pneumoconiosis, bronchitis, and emphysema in exposed *community*. Coal workers' pneumoconiosis (CWP) is characterized by development of coal macules, a focal collection of coal dust particles with a little reticulin and collagen accumulation. These lesions may be visible as small opacities (less than 1 cm in diameter) on X-rays. Complicated CWP is characterized by lesions consisting of a mass of rubbery well-defined black tissue that is often adherent to the chest wall. This is associated with decrements in ventilatory capacity, low diffusing capacity, abnormalities of gas exchange, low arterial oxygen tension, pulmonary hypertension, and premature death. The disease may progress after the cessation of exposure.

Effects on plant: Plants exposed to toxic dust exhibit lesser growth, become more prone to disease attack and rodent attack, their stomata and other holes get chocked, and they also gasp for oxygen. Fruits, vegetables, and cereals from these plants contain toxins.

Effects on infrastructure: Biochemical reactions occur when constituents present in the coal dust starts reacting with the constituents of the host surface materials, resulting in increase in the rate of weathering/weakening.

Pollution of Water

The potential impacts of mining on the water [28] are:

- Disruption of hydrological pathways
- Seepage of contaminated leachate into aquifers
- Depression of the water table around the dewatered zone
- Disposal of saline mine water into rivers

The impacts of mining arising from the disruption of hydrological pathways, seepage of contaminated leachate into aquifers, and depression of the water table tend to be relatively localized and limited compared to disposal of mine water. Disposal of mine water is a worldwide problem, occurring wherever operating mines, both underground and opencast, are found [19]. The quality of the mine water depends largely on the chemical properties of the geological materials that come in contact with the water.

The mining industry is a major producer of acidic sulfur-rich water that typically poses a risk to the environment. Mining increases the exposed surface area of sulfur-bearing rocks allowing for excess acid generation beyond natural buffering capabilities found in host rock and water resources. Pyrite (FeS_2) is the major sulfur mineral in coal. When coal is mined, fresh sulfur-bearing minerals in the coal and rocks are exposed to air and water. Problematic mine drainage is given many names including *acid rock drainage* (ARD), *acid mine drainage* (AMD), and mining influenced water (MIW). AMD forms during metal or coal mining when sulfur-bearing minerals are exposed to water and air, forming sulfuric acid. Heavy metals, leached from rocks, can combine with the acid and dissolve, creating highly toxic runoff. The general distinction between ARD and AMD depends on whether drainage quality has been degraded by mining or is of poor quality due, in part, to natural causes. MIW is the term used to refer all mining-related water because acidic, neutral, and alkaline water can all transport metals and other contaminants.

The origin of acidic metal-rich mine drainage water is the accelerated oxidation of FeS_2 and other sulfidic minerals. The reaction of pyrite with oxygen and water produces a solution of ferrous sulfate and sulfuric acid. Ferrous iron can further be oxidized producing additional acidity. The following reaction shows the reaction of pyrite with oxygen and water to produce hydrogen ions, sulfate ions, and soluble metal ions.

$$2FeS_2(s) + 7O_2 + 2H_2O \rightarrow 2Fe^{+2} + 4SO_4{}^{-2} + 4H^+$$

Further oxidation of Fe^{+2} (ferrous) to Fe^{+3} (ferric) occurs when sufficient oxygen is dissolved in the water or when the water is exposed to sufficient atmospheric oxygen.

$$2Fe^{+2} + {}^1\!/_2O_2 + 2H^+ \rightarrow 2Fe^{+3} + H_2O$$

Fe^{+3} can either precipitate as $Fe(OH)_3$, a red-orange precipitate seen in waters affected by ARD, or it can react directly with pyrite to produce more Fe^{+2} and acidity as per the following reactions:

$$2Fe^{+3} + 6H_2O \leftrightarrow 2Fe(OH)_3(s) + 6H^+$$

$$14Fe^{+3} + FeS_2(s) + 8H_2O \rightarrow 2SO_4{}^{-2} + 15Fe^{+2} + 16H^+$$

In undisturbed natural systems, this oxidation process occurs at slow rates over geologic timescale. Iron- and sulfur-oxidizing bacteria are also known to catalyze these reactions at low pH, thereby increasing the rate of reaction by several orders of magnitude [18]. Bacteria such as *Thiobacillus ferrooxidans* and *Ferroplasma acidarmanus* are known to specifically accelerate such reactions.

A few management options for saline mine water, as summarized by Annandale et al. [1], are:

1. Pollution prevention at source
2. Reuse and recycling of water to minimize the volume of polluted water
3. Treatment of effluents, if the problem cannot be solved through prevention, reuse, and recycling
4. Discharge of treated effluent
5. Utilization of gypsiferous mine water for irrigation

The term “gypsiferous” refers to rocks and soils containing more than 2% *gypsum*, i.e., calcium sulfate ($CaSO_4 \bullet 2H_2O$). There are concerns regarding potential use of mine water for agricultural crops [5], due to the amount of salt that would leach and potentially contaminate groundwater. The next few sections highlight successes and limitations with respect to environmental impact and sustainable use of such waters for irrigation of agricultural crops.

Restoration Process

It is the obligation of the mining company to restore the working cavity in the best possible way, through overburden disposal activities and technical reclamation, to the existing natural environment in terms of functionality and aesthetics. The following measures can be applied for the recultivation of deteriorated surfaces:

- Technical measures – to contribute to the improvement of resistant and deformable characteristics of the dump and directly influence the enhancement of the stability of slopes.

- Bio-technical measures – to contribute to faster achievement and maintenance of the dump stability along with the technical measures.
- Biological measures – that imply the implementation of agricultural and forest improvements that contribute to the stability and maintenance of reclaimed areas, but they are much more significant from the aspect of area revitalization and establishing natural biocenoses (a group of interacting organisms that live in a particular habitat and form a self-regulating ecological community)

The process of reverting an industrial land back to an agricultural terrain is a very slow process and good planning steps are required. Depending on the size of the coal deposit, the process of turning an agricultural, forest, or urban environment into an industrial (mining) one, and then, by the recultivation process, back again into the agricultural or forest one, may require long time, even many decades.

There are ongoing arguments in favor of preventing further expansion of the utilization of fossil fuels, such as coal. One of the restorative measures is to perform the recultivation of a dump even though it has a relatively low impact on the environment. Talking about fertile alluvial lands, it is necessary to preserve the fertile solum (a set of related soil horizons that share the same cycle of pedogenic processes) through selective excavation in order to bring the soil back to agricultural production or other use. The forestation of soil and terrains deteriorated by opencast lignite mining will prevent the further deterioration processes, contribute to the maintenance of the ecological balance in nature, and enhance the absorption of CO_2 from the air and increases the content of the oxygen therein.

The control of particulate emission is a fundamental part of a mine environmental management plan because of the increasing public awareness of human health issues and expectations of environmental performance, and the duty of care required of mine operators by government and the community. Particulate emission management system can result in cost savings, increased profits, and improved government and community relations, as well as easier access to resources and financial support in the future.

There are a number of systems to address emissions of harmful air pollutants from coal mines. It is technologically feasible to capture or flare methane from coal mines, instead of releasing methane directly into the atmosphere. Many mines have already taken steps toward capturing methane emissions for economic reasons. Twenty-three mines in Alabama, Colorado, Pennsylvania, Virginia, and West Virginia have methane drainage facility and they recover methane within the range of 3–88% efficiency [7]. Of these 23 mines, 12 sell recovered methane for natural gas energy use and two mines use the methane to heat mine ventilation air and to generate onsite power. Flaring is an option if capture is not technologically feasible. Although flaring occurs at a small number of mines in the USA, there is a long and safe history of flaring at working coal mines in the UK and Australia. US EPA estimates that nearly 50% of all of the US coal mine methane emissions, or more than 1.25 million tons of methane, can be reduced at a zero net cost, while nearly 90% can be reduced at a cost of less than $15/t. However, the benefit of reducing methane could be as much as $240/t of methane reduced. Any efforts to address methane will most likely also address VOC emissions due to the fact that VOCs and methane are often released together from coal mines.

Examples of particulate matter emission control measures currently in use include, but are not limited to: (a) storing coal in enclosed coal silos or barns; (b) paving coal mine access roads; (c) watering or treating with dust suppressant any unpaved roads; (d) enclosing conveyor transfer points; (e) use of dust collection baghouses or other controls to reduce emissions from transfer points and crushers within processing plants; (f) fitting out-of-pit conveyors with hoods or otherwise containing emissions; (g) fitting out-of-pit dump hoppers with water sprays, a baghouse, or other controls; (h) treating haul roads with dust control chemicals or water; (i) watering short-term haul roads; and (j) regularly maintaining haul roads to reduce dust re-entrainment.

With regard to NO_x, mines have reduced emissions by as much as 75% through the use of borehole liners and changing their blasting agent blends. Other mitigation measures that may be effective at reducing NO_x include reduced blast sizes, changed composition of explosive agents, and changed placement of blasting agents. Reduction of impacts due to radioactivity consists of ventilation for underground mines, dust control, standard containment measures for solid waste, and standard precautions for discharges from processing.

Sustainability and Mining

The Earth's resources have played a vital role for human communities. Nowadays, the management of our natural resources has become an urgent issue at both national and international level. The extraction and use of natural resources are also causing environmental problems, which require urgent solutions. For mining operations, resource extraction not only has a massive impact on ecosystems, it also releases pollutants contained in the rock, and consumes large amounts of water and energy. The transportation of resources from remote areas requires an ecologically intensive transport infrastructure.

Due to the progressive exploitation of mineral deposits and the availability of new technologies, deposits with lower ore content are now being mined. This means that an increasing amount of ore and other non-usable material has to be extracted to produce the same amount of metal. For example, toward the beginning of the twentieth century, copper ore mined by US mineral industry consisted of about 2.5% of usable metal by weight; today the proportion has dropped to 0.51% [8]. This activity impacts even more severely on ecosystems and water resources and increases the volume of mining waste, resulting in even more radical changes to entire landscapes. The rising global demand for resources accelerates this trend. After extraction, the subsequent stages in the raw materials' life cycle entail further environmental pollution. Comparative analyses of industrial sectors show that the highly resource-intensive industries are associated with above-average levels of emissions of greenhouse gases and other pollutants.

Climate change and the overexploitation of natural resources are two sides of the same coin. Climate change will impact water supplies, exacerbating existing pressures on water resources caused by population and economic growth. Given the combination of these stressors, the sustainability of water resources in future decades is a concern in many parts of the world. The sustainability of water resources is defined as the maintenance of natural water resources in adequate quantity and with suitable quality for human use and for aquatic ecosystems. Human needs for resources, like water, land, continue to grow with increasing population, primarily for direct consumption, but also secondarily for energy production, and agricultural and mining activities.

The privilege of mining can be enjoyed in the best possible way, but there is the responsibility of looking out for the future generations, preserving some of the natural resources by utilizing reserves in compliance with the principles of the sustainable development. There is also a need to reclaim and to revive deteriorated surfaces resulting from mining.

The nine sustainable development challenges that the mining industries face are:

- Ensuring long-term viability of the mining industry
- Controlling, using, and managing the land
- Using minerals to assist with economic development
- Making a positive impact on local communities
- Managing the environmental impacts of mines
- Maximizing the use of minerals so as to reduce waste and inefficiency
- Giving stakeholders access to information to build trust and cooperation
- Managing the relationship between large corporates and small-scale mining companies
- Sector governance: clearly defining the roles, responsibilities, and instruments for change expected of all stakeholders.

Mining, Minerals and Sustainable Development (MMSD) North America has developed an approach to assess how a mining/mineral project or operation contributes to sustainability. The assessment considered tracking the record of mining and minerals in the past and its current contribution to and detraction from economic prosperity, human well-being, ecosystem health, and accountable decision-making. A set of seven questions were developed as a means of assessing whether the net contribution to sustainability over the term of a mining/mineral project or operation will be positive or negative and a way of discovering how current activities can be improved and aligned with the emerging concept of sustainability [12]. The seven-part numbering used by MMSD (see Table 1) has been intended as an aid to

Mining Industries and Their Sustainable Management. Table 1 Seven questions to sustainability – How to assess the contribution of mining and minerals activities

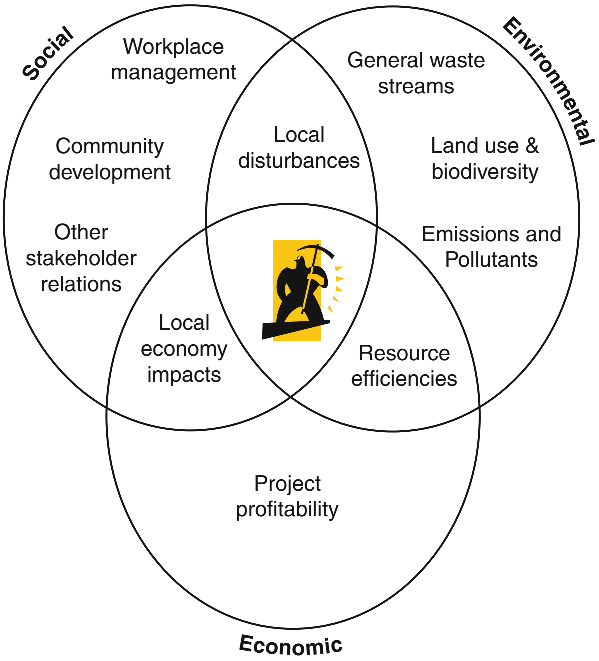
Seven fundamental questions	Framework for guiding a sustainability assessment
1. *Engagement* Are engagement processes in place and working effectively?	Processes of engagement are committed to, designed, and implemented so that they: ● Ensure all affected communities of interest have the opportunity to participate in the decisions that influence their own future ● Are understood, agreed upon by implicated communities of interest, and consistent with the legal, institutional, and cultural characteristics of the community and country where the project or operation is located
2. *People* Will people's well-being be maintained or improved?	Project/operation lead directly or indirectly to maintenance of people's well-being: ● During the life of the project/operation ● Post-closure
3. *Environment* Is the integrity of the environment assured over the long term?	Project or operation lead directly or indirectly to the maintenance or strengthening of the integrity of biophysical systems so that they can continue in post-closure to provide the needed support for the well-being of people and other life forms
4. *Economy* Is the economic viability of the project or operation assured, and will the economy of the community and beyond be better off as a result?	Assurance of the financial health of the project/company and contribution of the project or operation to the long-term viability of the local, regional, and global economy in ways that will help ensure sufficiency for all and provide specific opportunities for the less advantaged
5. *Traditional and nonmarket activities* Are traditional and nonmarket activities in the community and surrounding area accounted for in a way that is acceptable to the local people?	Project or operation contribute to the long-term viability of traditional and nonmarket activities in the implicated community and region
6. *Institutional arrangements and governance* Are rules, incentives, programs, and capacities in place to address project or operational consequences?	Institutional arrangements and systems of governance in place that can provide certainty and confidence that: ● Capacity of government, companies, communities, and residents to address project or operation consequences is in place or will be built ● Capacity will continue to evolve and exist through the full life cycle including post-closure
7. *Synthesis and continuous learning* Does a full synthesis show that the net result will be positive or negative in the long term, and will there be periodic reassessments?	Placement of an overall evaluation and periodic reevaluation based on: ● Consideration of all reasonable alternative configurations at the project level ● Consideration of all reasonable alternatives at the overarching strategic level for supplying the commodity and the services it provides for meeting society's needs ● Consideration of all reasonable alternatives at the overarching strategic level for supplying the commodity and the services it provides for meeting society's needs

communicate and not to imply a particular sequencing of steps or prioritization of topics. These questions follow a hierarchy of objectives, indicators, and specific metrics. In this way a single, initial motivating question cascades into progressively more detailed elements, which can be tailored to the project or operation being assessed for application throughout its full life cycle.

The focus on applying the seven questions to sustainability approach is not so much on how mining can be sustainable (mining as a discrete activity cannot continue indefinitely) but on how mining can contribute to sustainability [9].

Sustainability Opportunity and Threat Analysis is another simple operational tool addressing the social, economic, and environmental dimensions of the issues under consideration and it can be applied to evaluate the viability of a mining operation and its ability to contribute to sustainable development objectives. An impact-based model of sustainable mine development is shown in Fig. 2. This analysis involves constructing inventory of sources of impacts and following key steps including:

(a) Scoping (addressing the reasons for the mining process and the environment, and agreeing to the scope of the exercise)
(b) Information gathering (to emphasize the importance of collecting and organizing relevant information into a suitable framework)
(c) Identifying risks (systematically reviewing impact areas under consideration, and identifying opportunities or threats)
(d) Analyzing and evaluating the risks (qualitative scales of likelihood and consequence can be assigned to identify opportunity or threat to create overall risk rating and prioritized list)
(e) Treating risks (control measures to address opportunities or threats considered high priority)
(f) Reporting and reviewing to represent a broad, scanning exercise that can be picked up by existing business planning and monitoring processes. It also provides an opportunity for subsequent evaluation of relevant metrics for the issues identified by the process, allowing operations to measure both impacts and progress toward agreed objectives.



Mining Industries and Their Sustainable Management. Figure 2

Sustainable development model for impact-based mining operations

Most mine designs are based on traditional mining engineering factors, such as the quality of the commodity being mined, geology, topography, hydrology, land ownership, geography, infrastructure, etc. Environmental compliance and sustainability are to be considered in mine design and operation as a modifying factor to those designs. While practices may become more responsive to sustainability, mine design continues to be governed by established mining engineering approaches.

A review of the available literature on engineering optimization does not reveal any focus on mine design, environmental protection associated with mines, and sustainability. Mathematical multi-criteria optimization approaches, however, have been used in resource management. Mine design optimization would need to consider all constraints, system parameters and characteristics, and desired outcomes in order to build a useful and reliable model. Since optimization of mine design, and in particular coal mine design, to address sustainability along with other parameters has not been widely practiced, identifying the appropriate parameters for measurement and the mathematical or logical relationships between these parameters is not a trivial task.

Another important constraint on mining operations is the statutory and regulatory frameworks within which they are required to operate. Many of these legal and policy structures tend to create an adversarial approach by instituting a system where one or more parties must respond to actions, proposals, decisions, etc. of another party. The participation is often late in the design process, or after design has been completed, and thus any change to mine or reclamation design necessary as a result of public or regulatory agency input creates a retrofitted design, which cannot possibly be optimal. To optimize the design of mining and reclamation operations, traditional mining engineering considerations and environmental and sustainability goals must be accounted for simultaneously.

Several technologies are discussed that are considered to promote sustainable development.

Irrigation with Lime-Treated Acid Mine Drainage (AMD)

Higher crop yields were obtained under sprinkler irrigation with treated mine water compared to dryland production, without any foliar injury to the crop. Possible nutritional problems, for example deficiencies of potassium, magnesium, and nitrate, occurring due to calcium and sulfate dominating the system, can be solved through fertilization. Such soils need to be managed and fertilized differently to those on which crops are produced under normal farming conditions. Sugar beans, wheat, maize, and potatoes were very successfully produced under irrigation with calcium sulfate and magnesium sulfate-rich mine water. In an experimental setting, soil salinity increased with time, but the values of soil saturated electrical conductivity stabilized at relatively low levels, due to gypsum precipitation [1]. Measurements taken between 1997 and 2007 showed that soil salinity increased from a low base and oscillated around 250 mS m^{-1}.

Land preparation and fertilization management are, however, critical for successful crop production, especially on rehabilitated soil. During short to medium term (up to 8 years) irrigation with gypsiferous mine water, negligible impact was noticed by Annandale et al. [1] on groundwater quality. They operated the system with flexibility and managed with the multiple objectives like maximum crop production, water use, job creation, economic return or maximum gypsum precipitation, and minimum salt leaching. Gypsum precipitation was also shown to be taking place in the soil. The presence of gypsum did not create any physical and/or chemical property changes that could adversely affect crop production and soil management. Crop production under irrigation with coal-mine water, rich in calcium, magnesium, and sulfate is, therefore, feasible, and sustainable if properly managed.

Pasture production with sodium sulfate-rich mine effluent is also feasible, but requires a well-drained profile and a large leaching fraction to prevent unsustainable build up of salt in the soil. Unfortunately, this type of water does not present much of an opportunity for gypsum precipitation, which is able to drastically reduce the salt load of the receiving water in the case of calcium and sulfate-rich mine water. The application of calcium nitrate as a nitrogen source to the crop adds calcium to the soil and removes some sulfate from the water system by enhancing gypsum precipitation. Measurement of the hydraulic conductivity of the soil is recommended to monitor the effect of the water on the infiltration rate of the soil, as high sodium levels are likely to cause deflocculation or dispersion of clay particles.

Application of Phytotechnology

Phytotechnology can be applied to address issues related to stabilization of tailings and hydraulic control for drainage so that human and ecological exposures to contaminants associated with mining solid wastes and mine impacted waters are low. Implementation of phytotechnology is a common component of mining reclamation and *restoration* projects by the establishment of a plant cover as a final remedy. In certain cases, application of phytotechnology can be used for removal of metals from *contaminated media*. Establishing phytotechnology requires careful plant species selection and soil amendments that equates to an initial investment; however, these systems, once established can be maintained with minimal effort.

There are six basic phytoremediation mechanisms that can be used to clean up contaminated sites: phytosequestration, rhizodegradation, phytohydraulics, phytoextraction, phytodegradation, and phytovolatilization (see Table 2).

The particular phytotechnology mechanism used to address contaminants depends not only on the type of contaminant and the media affected, but also on the cleanup goals. Typical goals include containment through stabilization or sequestration, remediation through assimilation, reduction, detoxification, degradation, metabolization or mineralization, or both. Applying phytotechnology to impacted sites entails selecting, designing, installing, operating,

Mining Industries and Their Sustainable Management. Table 2 Phytotechnology mechanisms to sequester constituents of interest [13]

Mechanism	Description	Cleanup Goal
Phytosequestration	Ability of plants to sequester selected contaminants in the rhizosphere zone through exudation of phytochemicals and on the root through transport proteins and cellular processes	Containment
Rhizodegradation	Exuded phytochemicals can enhance microbial biodegradation of contaminants in the rhizosphere	*Remediation* by destruction
Phytohydraulics	Ability of plants to capture and evaporate water off the plant and take up and transpire water through the plant	Containment by controlling hydrology
Phytoextraction	Ability of plants to take up contaminants into the plant with the transpiration stream	Remediation by removal of plants
Phytodegradation	Ability of plants to take up and break down contaminants in the transpiration stream through internal enzymatic activity and photosynthetic oxidation/reduction	Remediation by destruction
Phytovolatilization	Ability of plants to take up, translocate, and subsequently transpire volatile contaminants in the transpiration stream	Remediation by removal through plants

maintaining, and monitoring planted systems that use the various mechanisms mentioned above. The goal of the system can be broadly based on the remedial objectives of containment, remediation, or both. Furthermore, the target media can be soil/sediment, surface water, or groundwater, and these can be either clean or impacted. In some cases, groundwater transitioning to surface water can be addressed as a riparian situation where target media are combined.

One of the main advantages of phytotechnology is that this technology can be applied to a variety of metals in mining sites and to impacted soil/sediment, surface water, and groundwater. In addition, it can be applied to various combinations of chemical types and impacted media simultaneously. Additional advantages are:

(a) Considered a green and sustainable technology
(b) Does not require supplemental energy, although monitoring equipment may use solar power
(c) Can improve the air quality and sequester greenhouse gases
(d) Minimal air emissions, water discharge, and secondary waste generation
(e) Lower maintenance, resilient, and self-repairing
(f) Inherently controls erosion, runoff, infiltration, and dust emissions
(g) Passive and in situ
(h) Favorable public perception
(i) Improves aesthetics, including reduced noise
(j) Applicable to remote locations, potentially without utility access
(k) Provides restoration and land reclamation during cleanup and upon completion
(l) Can be cost-competitive

The benefits of using the phytotechnology-based techniques are the relative lower costs, labor requirements, and safer operations compared to the more intensive and invasive conventional techniques. Phytotechnology generally provides long-term remedial solutions. Plantations may require irrigation, fertilization, weed control (mowing, mulching, or spraying), and pest control. Establishment of phytotechnology systems include various expenditures, such as earthwork, labor, planting stock, planting method, field equipment, heavy machinery (typically farming or forestry equipment), soil amendments, permits, water control infrastructure, utility infrastructure, fencing, security, etc. About 10–15% of the initial capital costs can be added as a contingency for replanting [13].

Phytotechnology are appropriate only under specific conditions. The major limitations are depth, area, and time. The physical constraints of depth and area depend on the plant species suitable to the site (i.e., root penetration) as well as the site layout and soil characteristics. Phytotechnology typically require larger tracts of land than many alternatives. Time can be a constraint since phytotechnology generally take longer than many other alternatives and are susceptible to seasonal and diurnal changes. Additional limitations include a plant's tolerance to specific constituents of concern or site conditions, availability of water as irrigation source, climate (challenging for plant establishment in areas with short growing season or in arid environments), and pests, infestations, or other nuisances.

Biomining and Bioprocessing

Mining industries are increasingly aware of the potential of microbiological approaches for recovering base and precious metals from low-grade ores, and for remediating acidic, metal-rich wastewaters that drain from both operating and abandoned mine sites. Biological systems offer a number of environmental and economical advantages over conventional approaches, such as pyrometallurgy, though the microbial application is not appropriate in every situation. *Biomining* (metal extraction) and bioprocessing are currently utilized in full-scale operations to process low-grade deposits and reprocessing earlier metal-containing wastes. This usually results in the production of less chemically active tailings, lower energy inputs, and other environmental benefits (zero production of noxious gases). Recently, there have been major advances in the field of microbiology, which will allow greater control of *bioleaching* operations, resulting in greater efficiencies and faster rates of metal extraction. Mineral processing using microorganisms have been exploited for extracting gold, copper, uranium, and cobalt. Engineering systems ranging from crude *heap leaching* systems to temperature-controlled bioreactors have been used, depending on the nature of the ore and the value of the metal product. A typical example of mineral bioprocessing mechanism is discussed below.

Sulfide minerals may be divided into acid soluble (such as zinc sulfide or sphalerite) and acid insoluble (such as pyrite and arsenopyrite). Two routes (the "thiosulfate" and "polythionate" mechanisms) have been proposed for the biological oxidation of these sulfide minerals [21]. Acid-soluble sulfides are readily degraded by sulfur-oxidizing *acidophiles*. The mineral is first subjected to proton-mediated dissolution, forming the free metal and hydrogen sulfide:

$$MS + 2H^+ \rightarrow M^{2+} + H_2S$$

The hydrogen sulfide so formed is microbially oxidized to sulfuric acid, allowing the process to continue:

$$H_2S + 2O_2 \rightarrow H_2SO_4$$

Acid-soluble sulfides may also be attacked by ferric iron, producing ferrous iron and polysulfide:

$$MS + Fe^{3+} + H^+ \rightarrow M^{2+} + 0.5H_2S_n + Fe^{2+}$$

Polysulfide may be further oxidized by ferric iron to produce elemental sulfur, which, in turn, is oxidized to sulfuric acid, and will further accelerate mineral dissolution via proton attack:

$$0.5H_2S_n + Fe^{3+} \rightarrow 0.125S_8 + Fe^{2+} + H^+$$

$$0.125S_8 + 1.5O_2 + H_2O \rightarrow 2H^+ + SO_4{}^{2-}$$

Sulfides that are resistant to proton attack are oxidized by ferric iron, producing thiosulfate as an initial by-product:

$$FeS_2 + 6Fe^{3+} + 3H_2O \rightarrow 7Fe^{2+} + S_2O_3{}^{2-} + 6H^+$$

Accelerated oxidation can result in low pH, high concentrations of dissolved metals and, in some cases, elevated temperatures. These conditions limit the diversity of life-forms that occurs in commercial bioleaching operations. Single-celled organisms live only in extremely acidic liquors (pH <1–4) and are obligate acidophiles, some are thermophilic (to varying degrees) and some can fix carbon dioxide. The primary microorganisms involved in mineral oxidation are those that catalyze the oxidation of ferrous iron and/or reduce sulfur, while others contribute to the process indirectly by, for example, removing materials that accumulate during ore dissolution [14].

Engineering approaches used in biomining are: (a) irrigation-based principles (dump- and heap-leaching, and in situ leaching) and (b) stirred tank processes [20]. The dump leaching involves gathering low-grade copper-containing ore of large rock/boulder size into vast mounds or dumps and irrigating these with dilute sulfuric acid to encourage the growth and activities of mineral-oxidizing acidophiles, primarily iron-oxidizing *mesophiles.* Copper can be precipitated from the metal-rich streams draining from the dumps by displacement with iron. Other developments on the engineering and hydrometallurgical aspects of biomining have involved the use of thin layer heaps of refractory sulfidic ores (mostly copper, but also gold-bearing material) stacked onto water-proof membranes, and recovery of solubilized copper using solvent extraction coupled with *electrowinning*. In situ bioleaching has been developed to scavenge for uranium and copper in otherwise worked out mines.

Bioremediation

Water draining from active and abandoned mines and mine wastes are acidic and often contain elevated concentrations of metals (iron, aluminum and manganese, and others) and metalloids (arsenic, chromium, selenium, and uranium). The basis of bioremediation of AMD derives from the abilities of some microorganisms to generate alkalinity and immobilize metals, thereby essentially reversing the reactions responsible for AMD.

Microbiological processes that generate net alkalinity are mostly reductive processes, and include *denitrification*, *methanogenesis*, sulfate reduction, and iron and manganese reduction. *Ammonification* (the production of ammonium ion from nitrogen-containing organic compounds) is also an alkali-generating process. Photosynthetic microorganisms, by consuming a weak base (bicarbonate) and producing a strong base (hydroxyl ions), also generate net alkalinity:

$$6HCO_3^-(aq) + 6H_2O \rightarrow C_6H_{12}O_6 + 6O_2 + 6OH^-$$

The reduction of soluble iron (ferric iron) does not decrease solution acidity, however, the reduction of solid phase (*crystalline* and *amorphous*) ferric iron compounds does.

$$Fe(OH)_3 + 3H^+ + e^- \rightarrow Fe^{2+} + 3H_2O$$

where e^- represents an electron donor, which is generally an organic substrate.

Bacteria that catalyze the *dissimilatory* reduction of sulfate to sulfide generate alkalinity by transforming a strong acid (sulfuric) into a relatively weak acid (hydrogen sulfide).

$$SO_4^{2-} + 2CH_2O + 2H^+ \rightarrow H_2S + 2H_2CO_3$$

Besides the ameliorative effect on AMD brought about by the resulting increase in pH, the reduction of sulfate is an important mechanism for removing toxic metals from AMD, since these metals (e.g., zinc, copper, and cadmium) form highly insoluble sulfides.

$$Zn^{2+} + H_2S \rightarrow ZnS + 2H^+$$

The bioremediation technologies, like constructed wetlands and compost bioreactors for AMD, are passive systems. In case of the constructed wetlands, the solid-phase products of water treatment are contained within the wetland sediments. The key major advantages of these passive bioremediation systems are their relatively low maintenance costs. The limitations of these technologies are:

(a) They are often relatively expensive to install.
(b) May require more land area than is available or suitable.
(c) Their performance is less predictable than chemical treatment systems.
(d) The long-term fate and stability (in the case of compost bioreactors) of the deposits that accumulate within them is uncertain [15].

Aerobic wetlands are generally more effective to treat mine waters that are net alkaline. The reaction involves oxidation of ferrous iron, and subsequent hydrolysis of the ferric iron produced, which is a net acid-generating reaction.

$$4Fe^{2+} + O_2 + 4H^+ \rightarrow 4Fe^{3+} + 2H_2O$$

$$4Fe^{3+} + 4H_2O \rightarrow 4Fe(OH)_3 + 12H^+$$

In the event the pH of the mine water decreases significantly as a result of these reactions, additional amendments (such as an anoxic limestone drain) may be required. To maintain oxidizing conditions, aerobic wetlands are relatively shallow systems that operate by surface flow. Though *macrophytes* are planted for aesthetic reasons, they also regulate water flow, accelerate the rate of ferrous iron oxidation, and stabilize the accumulating iron precipitates.

In contrast to aerobic wetlands, the compost bioreactors (installations that are enclosed entirely below ground level and do not support any macrophytes) are anaerobic. The microbially catalyzed reactions in compost bioreactors generate net alkalinity and biogenic sulfide, and thus can treat mine waters that are acidic and metal-rich. The reductive reactions that occur within compost wetlands are driven by electron donors that derive from the organic matrix of the compost itself. The indigenous iron- and sulfate-reducing bacteria are generally considered to have the major roles in AMD remediation in compost bioreactors.

Metal Recovery: Biotic and Abiotic

Innovative and alternative techniques that facilitate the economic control and recovery of metal values are beneficial not only for protection of human health and the environment, but also for the recovery of these commodities and resource conservation. AMD, rich in metals, should be considered not only as a serious environmental problem, but as an important resource of metals of considerable value to many industrial concerns.

Recently developed metal recovery processes from AMD and other acidic streams enable us to recover and recycle these metals cost effectively. A two-stage process was developed by Tabak et al. [23] to separate the sulfate-reducing bacteria (SRB) from AMD via sulfate reduction and production of biogenic hydrogen sulfide from the four-stage and six-stage metal sequential separation and biorecovery units. The four-stage selective sequential batch-type metal precipitation process was able to separate metal sulfides and hydroxides at reasonably high recovery rate and at high precipitate purities. Copper sulfide and zinc sulfide were precipitated in Stage 1, aluminum hydroxide was precipitated in Stage 2, ferrous sulfide was precipitated in Stage 3, and other metals were precipitated in Stage 4. The effluent water contained only calcium and magnesium in trace concentrations and both sulfate and metal sulfide concentrations were below detectable limits.

von Fahnestock [30] recently developed a breakthrough technology to efficiently remove sulfate ions and metal cations from AMD by using the Acid Mine Drainage Value Extraction Process (AMD VEP), which is a novel adaptation of liquid–liquid extraction process. The process converts AMD to purified water as well as saleable products, such as potassium sulfate and iron sulfate. This technology results in simultaneous cost-effective isolation and concentration of useful metals and sulfate ions from mine pool water. A 30-gpm (113.5-L/m) demonstration plant was built and was operational within 15 months in St. Michael, PA. In this technology, AMD water, laden with sulfate and iron, feeds into the water purification stages where the AMD water is sequentially contacted with an *extractant* solution in a countercurrent flow path. The extractant solution is formulated to efficiently pull the sulfate and iron from the aqueous phase.

The water purification stages are composed as a set of four mixer-settler units, which are two-compartment tanks. The different stages are described below:

1. In the first compartment, the extractant and the AMD are mixed in a chamber. The residence time is between 60 and 90 s.
2. In the second stage, the combined effluent flows from the mixing chamber into the settling chamber where the organic extractant phase disengages from the water phase. The extractant, containing iron and sulfate, overflows an exit weir into the settling chamber, and is separated cleanly from the water phase, which underflows the same weir and exits as a separate stream with proportionately less iron and sulfate.
3. In the recovery stages of the AMD VEP, the extractant, loaded with iron and sulfate, flows sequentially from the water purification stages to the metals recovery and sulfate recovery stages of the process. Iron and sulfate are sequentially recovered as usable products during these stages. The metals-recovery stages are a set of two to three metal cation recovery tanks. The extractant flows counterclockwise with an aqueous sulfuric acid solution to form an iron sulfate concentrate that is harvested for reuse. The sulfuric acid and the extractant are mixed and separated in a similar manner to that in the mixer-settler tanks of the extraction section.
4. The sulfate-recovery stages remove sulfate from the extractant, which again flows through a series of

mixer-settler tanks in a countercurrent fashion. The extractant is contacted with potassium carbonate (a basic aqueous solution) to produce a potassium sulfate concentrate. The extractant exits the last mixer settler of the sulfate-recovery section regenerated and ready to contact a new stream of AMD feedwater in the water purification stages. The potassium sulfate (K_2SO_4) concentrate leaves the sulfate-recovery stages and is collected and stored for sale as fertilizer or reuse in a product tank.

An engineering analysis indicated that the maximum cost to treat and recover metals by AMD VEP process is $8.00 dollars/1,000 gal [3] dependant on product values and reagent pricing. AMD VEP, based on abiotic processes, is one of the few recent technologies to purify AMD water sufficiently to enable its discharge to surface waters. In addition, the product water can be made useful for industrial, municipal, agricultural, residential, and other uses. Importantly, these process products enable the system to focus on meeting seasonal and regional demands. For example, potassium sulfate fertilizer can be produced for 6 months a year for the turf grass industry, and the same system can produce sodium sulfate deicer for highways for 6 months a year, by changing operating conditions and reagent feeds.

Water Quality and Acid Mine Drainage: Pre-mine Predictions and Post-mine Comparisons

Necessity of accurate prediction of acidic drainage from proposed mines is recognized by both industry and government as a critical requirement of mine permitting long-term operation. Substantial emphasis has been placed on prediction of acid drainage associated with coal development in the Eastern USA [2], and metal mining in the Western USA and in Canada [16]. The prediction of acid-generating potential from any geologic formation is dependent on the ability to characterize the presence and quantity of both acid-forming minerals and neutralizing minerals in the materials that are expected to be unearthed during mining operations. Typically, samples are collected by drilling during exploration, analyzed and interpreted with respect to their risk of acid formation. In these analyses, the amount of sulfur present in geologic materials is measured and attributed to being either an acid-forming mineral such as pyrite (FeS_2) or non-acid-forming mineral such as gypsum ($CaSO_4 \bullet 2H_2O$). The relative amount of acid-forming minerals is then contrasted to the amount of neutralizing minerals such as calcite ($CaCO_3$) to develop a prediction of the probability of acid generation.

The acid base accounting (ABA) of a material is the balance between total acid-generating potential (AP), which is the total amount of acidity that would be produced if all sulfide in a material is completely oxidized, and total acid-neutralizing potential (NP), which is the amount of acid that could be consumed by neutralizing minerals. AP and NP are converted to $CaCO_3$ equivalents and reported as grams of $CaCO_3$ per kilogram rock. ABA is typically calculated from analysis of sulfide S and carbonate C, assuming a 1:1 molar ratio of sulfide S (AP) and carbonate C (NP).

Converting chemical analysis for sulfide S (S_{FeS_2}) and carbonate C (C_{CaCO_3}) can be expressed as [26]:

$$AP = S_{FeS_2}(10) \times (3.12)$$

$$NP = C_{CaCO_3} \times (10) \times (8.33)$$

where

S_{FeS_2} = concentration sulfide sulfur in sample (weight% S)

3.12 = molecular weight of $CaCO_3$/molecular weight of sulfur

C_{CaCO_3} = concentration carbonate carbon in sample (wt% C)

8.33 = molecular weight of $CaCO_3$/molecular weight of carbon

The ratio of neutralization potential (NP) to acid potential (AP) is commonly presented in graphical interpretations with the inference that geologic materials with an abundance of NP are unlikely to generate acidic drainage. Skousen et al. [22] reported that NP:AP ratios <1 commonly produce acidic drainage, NP:AP ratios between 1 and 2 may produce either acidic or neutral drainage, and NP:AP ratios >2 should produce alkaline water.

However, this index does not always accurately predict the resultant acid generation from a mine. Out of 56 mines evaluated by Skousen et al. [22], 11% did not conform to the expected results based on NP: AP ratios, including four sites with ratios >2 that eventually produced acidic drainage. Mineralogical variation between each geologic domain causes dissimilar reactivity to weathering conditions and leads to laboratory variability in assessment. Forecasting future water quality impacts from AMD based on laboratory and field data should not be considered routine and robust, rather they should be considered an area of uncertainty and ongoing research.

Future Directions

The starting point is typically the acceptance of sustainable developmental principles at board-room level as corporate goals, and then informing the workforce, investors, and others of that commitment. Relevant employees need to be engaged as a first step in the practical application of sustainable development principles, followed by the gradual extension of training in sustainable methods of working to the workforce as a whole.

Bibliography

1. Annandale JG, Beletse YG, Stirzaker RJ, Bristow KL, Aken ME (2009) Is irrigation with coal-mine water sustainable? In: Proceedings of international mine water conference, Pretoria, South Africa
2. Brady KBC, Hornberger RJ, Fleeger G (1998) Influence of geology on postmining water quality: Northern Appalachian Basin. In: Brady, KBC, Smith MW, Schueck J (eds) Coal Mine Drainage Prediction and Pollution Prevention in Pennsylvania: Harrisburg, Pennsylvania. Department of Environmental Protection, p. 8–1 to 8–92
3. Conkle HN (2008) Engineering analysis report, value recovery from mine or rock drainage water, with water purity enhancement, using innovative low cost technology. Battelle, Columbus
4. DEFRA (2007) Waste strategy for England 2007. Presented to parliament by command of her majesty. Department for Environment, Food and Rural Affairs, PB12596
5. Du Plessis HM (1983) Using lime treated acid mine water for irrigation. Water Sci Technol 15:145–154
6. EIA (2010) Annual energy outlook 2010 with projections to 2035. U.S. Energy Information Administration, Office of Integrated Analysis and Forecasting, U.S. Department of Energy, Washington, DC. DOE/EIA-0383
7. Earth Justice (2010) Petition for rulemaking under the clean air act to list coal mines as a source category and to regulate methane and other harmful air emissions from coal mining facilities under section 111. Denver, CO
8. Gardner G, Sampat P (1998) Mind over matter: recasting the role of materials in our lives. Worldwatch Institute, Washington, DC, p 18
9. Hodge RA (2004) Mining's seven questions to sustainability: from mitigating impacts to encouraging contribution. Episodes 27(3):177–184
10. IEA (2007a) World energy outlook 2007 – China and India insights. OECD/IEA, Paris
11. IEA (2007b) CO_2 emissions from fuel combustion, 1971–2005, 2007 edition. OECD/IEA, Paris
12. IISD (2002) Seven questions to sustainability: how to assess the contribution of mining and minerals activities. Task 2 work group, mining, minerals and sustainable development North America (MMSD) North America. International Institute for Sustainable Development, Winnipeg
13. ITRC (2010) Technology overview of phytotechnologies. Prepared by The Interstate Technology & Regulatory Council, Mine Waste Team. Washington DC
14. Johnson DB (2001) Importance of microbial ecology in the development of new mineral technologies. Hydrometallurgy 59:147–158
15. Johnson DB, Hallberg KB (2002) Pitfalls of passive mine drainage. Rev Environ Biotechnol 1:335–343
16. MEND (2001) List of potential information requirements in metal leaching/acid rock drainage assessment and mitigation work. Mining environment neutral drainage program. W. A. Price, CANMET, Canada Centre for Mineral and Energy Technology
17. Mishra PC, Jha S (2010) Dust dispersion modeling in opencast coal mines and control of dispersion in Mahanadi coalfields of Orissa. Bioscan 2:479–500
18. Nordstrom DK, Southam G (1997) Geomicrobiology – interactions between microbes and minerals. Mineral Soc Am 35:261–390
19. Pulles W (2006) Management options for mine water drainage in South Africa. In: WISA, mine water division (ed) Mine water drainage-South African perspective. Water Institute Southern Africa, Johannesburg
20. Rawlings DE (2002) Heavy metal mining using microbes. Annu Rev Microbiol 56:65–91
21. Schippers A, Sand W (1999) Bacterial leaching of metal sulfides proceeds by two indirect mechanisms via thiosulfate or via polysulfides and sulfur. Appl Environ Microbiol 65:319–321
22. Skousen J, Simmons J, McDonald LM, Ziemkiewicz P (2002) Acid-base accounting to predict post-mining drainage quality on surface mines. J Environ Qual 31:2034–2044
23. Tabak HH, Scharp R, Burckle J, Kawahara FK, Govind R (2003) Advances in biotreatment of acid mine drainage and biorecovery of metals: 1. Metal precipitation for recovery and recycle. Biodegradation 14:423–436
24. U.S. EPA (1995) Identification and description of mineral processing sectors and waste streams. Office of Solid Waste, Washington, DC. RCRA Docket No. F-96-PH4A-S0001
25. U.S. EPA (2006a) Global mitigation of non-CO_2 greenhouse gases. Office of Atmospheric Programs, Washington, DC. EPA 430-R-06-005
26. U.S. EPA (2006b) Management and treatment of water from hard rock mines. Engineering Issue. Office of Research and

Development National Risk Management, Research Laboratory Cincinnati, OH. EPA/625/R-06/014

27. U.S. EPA (2009) Identifying opportunities for methane recovery at U.S. coal mines: profiles of selected gassy underground coal mines 2002–2006. Coalbed Methane Outreach Program, EPA 430-K-04-003
28. Younger PL, Wolkersdorfer C (2004) Mining impacts on the fresh water environments: technical and managerial guidelines for catchment scale management. Mine Water Environ 23:S2–S80
29. SDWF (2011) Mining and water pollution. Available from safe drinking water foundation. http://www.safewater.org/PDFS/resourcesknowthefacts/Mining+and+Water+Pollution.pdf. Accessed 11 May 2001
30. von Fahnestock M (2010) AMD value extraction process. Water Cond Purificat 52(1):38–40
31. Waddell S, Pruitt B (2005) The role of coal in climate change: a dialogic change process analysis. The generative dialogue project launch meeting, New York, NY, 6–8 Oct 2005
32. WEO (2007) World energy outlook 2OO7: China and India insights. International Energy Agency. 61 2007 01 1 P1, Paris

Mining Solid Wastes

JAAK J. K. DAEMEN[1], HALUK AKGÜN[2]
[1]Department of Mining Engineering, University of Nevada, Reno, NV, USA
[2]Department of Geological Engineering, Middle East Technical University (METU), Ankara, Turkey

Article Outline

Glossary

Acid drainage Acidic water ($pH < 7$, but frequently < 5) released from waste dumps, tailings dams, formed as a result of reactions between inflowing water, e.g., rainwater, and acid-generating minerals, e.g., sulfides, in particular pyrite, present in the dumps. One of the major environmental hazards associated with some types of mining.

Beneficiation Processing of ore, e.g., through crushing, grinding, gravity separation, flotation, in order to recover metals of economic value.

Coal washing Treatment of raw coal to remove noncombustible rocks such as shale and sandstone and to produce clean, washed, coal.

Heap leaching Recovering metal values by leaching the metal out of broken rocks stacked in heap leach dumps.

Reclamation The process of restoring a site, e.g., mine waste dumps, to an acceptable condition for future use, e.g., by controlling and minimizing any environmental impacts the waste might have.

Stripping ratio Ratio of barren overburden ("waste") rock that needs to be removed in a surface mine, e.g., an open pit mine or a strip mine, in order to reach the mineral of economic value (e.g., metal ore, coal seam, limestone bed).

Tailings Leftover very fine particles (typically size ranges <0.1 mm) from mineral processing operations, the recovery of minerals of value from ore after the ore has been crushed and finely ground.

Tailings dams Dams, earth structures, containment embankments, built up from tailings.

Tailings ponds Lagoons in which tailings are dumped, and allowed to settle, after which the water is recovered, preferably recycled.

Waste dumps Earth fill like dams built of broken blasted rock, e.g., overburden, barren rock overlying ore deposits of value, that needs to be removed in order to gain access to the ore deposit.

Definition of the Subject

Miners dig holes in the ground to excavate materials that contain minerals of value to and needed by society. An unavoidable by-product of such mining is the creation of vast amounts of solid wastes. Unavoidable, because the minerals of interest, e.g., copper, platinum, and gold, typically occur in extremely small concentrations, even in so-called rich ores. One major type of solid waste associated with mining is the barren rock that needs to be removed to gain access to any mineral deposit of commercial interest. A second major type of solid waste created by mining is the solids residue that results from the processing of the ore in order to extract the minerals of economic value. Usually ores need to be broken up, crushed, into relatively small particles, and sometimes need to be ground into even smaller, minute fragments in order to allow the recovery of the minerals of interest. The overwhelming majority of these small particles need to be disposed of as waste, usually referred to as tailings.

In light of the huge volumes, tonnages, generated by mining, solid waste management constitutes a major challenge for mining and for the engineers who design, plan, operate, and close mines. Not only are the amounts of solids vast, but sometimes they are potentially harmful to humans and to the environment. It is an essential and critical responsibility for modern mining to ensure that solid wastes are managed satisfactorily, i.e., with minimal and not more than acceptable detrimental impact to people and to the environment. What is acceptable to people has changed dramatically in recent decades, and as a result the control requirements on solid wastes generated by mining are numerous, stringent, and complex.

Mine wastes have been generated for millennia, although the amounts of waste generated in the distant past were trivial compared to the volumes generated in recent decades, and in the foreseeable future. Mine waste dumps remaining from ancient times now have become archaeological sites of great value for the study of ancient civilizations, and notably for the study of the early development of the use of a variety of metals, as well as of rock, whether for tools (flint, obsidian) or building stone or monument and statue construction. However, some of these ancient sites also have become and remain sources of pollution, notably acid rock drainage, indicating how long such long problems can persist if not treated or addressed adequately.

Although destructive consequences of solid mine waste have been recognized for at least several centuries, it is primarily since the 1970s that a more complete understanding has developed of the environmental and health problems that may be associated with or result from such wastes. By now the seriousness of some of the problems induced by solid mine wastes have become more fully appreciated and more widely recognized. As a result, stringent, and ever more stringent, regulations have been introduced for the management and control of mine wastes. Many mining companies have committed to comply with rigorous homeland regulations even in locations where such regulations may not yet have been legally or formally implemented. In parallel with the regulatory development has been a massive research development aimed at improving the understanding of environmental problems caused by solid mine wastes, and at improving practices for preventing, minimizing, and controlling such problems. As a result, a vast literature now is available, dealing extensively and in detail with virtually any aspect one can think of environmental degradation that might result from solid mine wastes. A major challenge that remains, however, in many parts of the world, is how to deal with a legacy of abandoned mine sites, usually accompanied by mine waste dumps, remaining from the past. In the USA, it has been estimated, although with considerable uncertainty, that such sites, in 12 western states, may number 33,000 ([1], who also provides references to more detailed reports). An obvious prime difficulty in dealing with such sites is that usually no ownership, no responsible party, can be identified, or exists anymore, and hence the responsibility for cleanup and site restoration defaults to society at large, often at considerable cost.

Introduction

Mining provides the raw materials such as base, ferrous, and precious metals; construction materials such as sand, gravel, building stone, and crushed rock; and industrial minerals such as salt, fertilizer (e.g., phosphate), clay (e.g., for ceramics), diatomaceous earth (e.g., for filter applications in the food and beverage industries), and energy, e.g., coal, lignite (brown coal), uranium, on which society critically depends. Associated with mining, unavoidably and intrinsically, is the generation of large volumes and tonnages of waste. Mining generates waste in a variety of ways. For surface mining, e.g., open pit mines, strip mines, or quarries, it usually is necessary to remove overburden, i.e., soil and rock overlying the deposit to be mined, in order to gain access to the mineral deposit to be mined. Depending on the local geology, this may involve the removal of materials ranging in thickness from meters to tens of meters, occasionally up to hundreds of meters. Overburden removal frequently constitutes a large fraction, often the largest fraction, of solid "waste" associated with surface mining.

A second source of solid wastes is the products that result from the processing of the ore. A fairly extreme and obvious example: a typical gold deposit may contain a few grams of gold per ton of ore. In order to liberate the gold from the ore, the ore needs to be broken up into small particles. Virtually all the crushed and ground rock will be returned to "waste" piles or dumps, except for the minute fraction of gold removed from the ore. While the metal content of base metals may be somewhat higher, e.g., 0.5–1% for copper, that still leaves typically more than 90% and frequently more like 99% of the processed ore to be disposed of as waste. Similarly, inert material removed from coal, in coal washing plants, can generate waste fractions of the order of 5–25% and more of the coal mined.

Given the large quantities of material mined, the volumes and tonnages of waste associated with mining are very large indeed. For reasons briefly touched on in Section Future Directions, it seems highly unlikely that these volumes can be decreased, at least not in the near future – if anything, it is virtually certain that they will increase, probably substantially.

Depending on the methods by which waste is generated, and depending on the controls required to assure that waste does not pose unacceptable hazards to the environment and to people, a variety of methods are used to dispose of solid wastes generated by mining. In strip mining of coal and lignite, by far the major surface mining method of these materials, the mining operations are designed and planned such that the overburden materials, the soil and rock overlying the beds to be mined, are emplaced in previously mined pits or strips. With these types of operations the overburden waste essentially is used and emplaced such that the original, pre-mining, surface contours and topography are restored, more or less, approximately, within a relatively short time after mining, and that reclamation and site restoration can progress simultaneously with, concurrent with, ongoing mining.

In open pit mining, e.g., for precious and base metals, the overburden material typically is emplaced in waste dumps, disposal piles virtually always located at a very short distance from the mine. Depending notably on the geochemical and mineralogical composition of the overburden, such waste piles or dumps may require more or less intensive environmental control measures.

Tailings, finely ground rock particles that result from mineral processing of ores, will require particular care for disposal. Typically, usually, they will be emplaced in tailings ponds and tailings dams. Usually, virtually always, just as for coal slimes, residue from coal washing plants, these types of waste will pose potentially serious environmental risks and hazards. Hence it is imperative that these structures be planned, designed, built, operated, and closed with particular care and with great emphasis on ensuring that any environmental impacts be kept to acceptable levels.

Of particular concern for tailings, and for some types of overburden wastes, is the risk that the materials might contain minerals, most likely sulfides, that, when allowed to react with water and oxygen, might result in the generation of acid releases. Historically, almost certainly, this has constituted the major environmental impact from many mining operations. It is essential, for modern mining operations, to assure that such acid releases be minimized.

For some ores, health hazards are created by the materials mined. This includes notably uranium tailings, often deposited and present in areas where uranium has been or is being mined. The uranium ore virtually always is milled, i.e., processed, near the mine. The milling produces finely ground uranium ore, and not all the uranium can be extracted from the ore. Hence the tailings, the residue, will contain some radioactive particles. Toxic elements such as mercury and arsenic frequently are associated with precious metal, e.g., gold, deposits and may require particular precautions in order to prevent unacceptable health risks associated with these materials, e.g., on waste dumps, in tailings ponds, on tailings dams, etc. Of particular concern in this regard may be dust generated by wind blowing over such structures, especially when dry, and if left without or with minimal surface protection.

Some types of wastes generated by mining have found useful uses, although this certainly is not common, notably as backfill for underground mines, in order to improve the stability of underground mine excavations, and in order to minimize impacts on the surface, e.g., in the form of surface subsidence, induced by underground mining. While such uses of solid mine waste certainly are important and significant in some mining districts or areas, they remain a relatively minor factor in overall solid mine waste management.

Whatever types of wastes are generated, and whatever waste management approaches are implemented, it is universally accepted good practice today, and a legal and regulatory requirement in most countries, to assure that mining operations, including wastes, be fully restored to an environmentally acceptable condition. Reclamation and site restoration is an integral part of today's mining planning and design practice. This does not imply or suggest that it is trivial or easy or not extremely costly, but responsible mining companies and operators recognize the need to meet society's needs for minerals using methods that are acceptable to society – their customers.

Types of Mining Solid Wastes

Solid mine wastes can be classified according to a number of different schemes, each of which has its own advantages and disadvantages, e.g., in terms of potential health and environmental impacts.

One basic yet important classification is by considering strictly whether or not the waste has a potential detrimental health or environmental impact. For example, a coarsely broken but relatively pure limestone, whether it be removed from above a coal deposit in order to gain access to the coal, or whether it be removed from above a marble deposit, in order to gain access to the marble, is extremely unlikely to, by itself, constitute either an environmental or a health hazard. (Admittedly, even if a relatively pure limestone was dumped in a river, it could have an environmental impact, e.g., resulting in flooding, or acting as an accidental dam.) Conversely, a shale bed, removed for similar purposes, that contains even a fairly small fraction of pyrite (iron sulfide) most certainly could constitute an environmental risk: oxidation of the pyrite, when in contact with water and oxygen, could generate an acidic effluent that could contaminate surface and ground waters.

Solid mine wastes can be classified on the basis of the particle size of the waste products. Both from an environmental control point of view and from a potential health impact point of view, the particle size and the particle size distribution, i.e., the particle grading, are solid waste characteristics of major importance. In addition, particle size and particle size distribution are dominant variables in designing disposal structures in which they are contained, because they are major engineering variables that will enter in stability and in hydrological, i.e., water flow, analyses. The particle size may range from exceedingly fine (e.g., phosphate slimes, <0.01 mm), to very fine, e.g., most tailings that result from metal ore processing, typically <0.1 mm, to very coarse, e.g., typical blasted overburden, where most of the particles will have a size exceeding many centimeters, and a large fraction of the particles will have sizes of the order of 1 m and more, sometimes considerably more. Once dried, dust conditions associated with these different particle size ranges will be very different. Coarse materials will allow ready and easy water drainage, and will not become airborne, even under extreme wind conditions. Fines are likely to be difficult to drain, are likely to have extremely low hydraulic conductivities, and may very easily be blown away by wind, sometimes to considerable distances.

From an environmental impact point of view, it almost certainly is the geochemical composition of the waste, and sometimes, in addition, the chemistry of any products that have been used in the beneficiation of the ore, that will be the major factor. Of prime concern are sulfides, and in particular sulfides that could result in the release of acid drainage, undoubtedly the major environmental problem associated with and generated by solid mine wastes. Equally important, whenever present, if present in sufficient concentration (which may be very low), are heavy metals, e.g., cadmium, selenium, mercury, chromium, and lead. These elements, generally toxic to humans, even in relatively small concentrations, also frequently are toxic to plants and animals. They are naturally present in many metallic orebodies, although in widely varying concentrations and chemical forms, and certainly their release needs to be controlled and limited to acceptable levels.

Solid mine wastes routinely are classified on the basis of the products mined. Hence a clear distinction is made between coal mine wastes, metal mine wastes, and industrial mineral mine wastes. Not surprisingly, the tailings left from milling uranium ores, and some other radioactive ores, are treated as a class of materials by themselves, given that they release radionuclides. Although such categorizations certainly are helpful, they also are somewhat simplistic, and do not allow for a full accounting for any potential impacts of the wastes: depending on the overburden characteristics, especially chemically, but also physically, coal mine wastes or metal mine wastes may pose significantly different threats or risks, even when resulting from mining the same product of ultimate interest and value.

Sources of Mining Solid Wastes

By far the largest source of solid mine waste is the rock that needs to be removed above minerals of value in order to reach these minerals and allow their recovery. A second major source of solid waste is the residues of the processing that is performed in order to recover the minerals of value. For coal, this frequently involves coal washing. In order to wash the coal, it usually will be ground to a relatively small size (also preferred from a combustion point of view, in burners used for power generation). Similarly in order to allow recovery of precious and base metal values, e.g., platinum, gold, silver, copper, lead, and zinc, it usually is necessary to grind the rock to a very small size (<1 mm, often <0.1 mm), because that is the particle size range in which the metals can be liberated. Iron ore, on the other hand usually does not require sizing to anywhere near such small sizes, whereas phosphate (a major fertilizer source) typically is ground to a much smaller size. In these extremely small size ranges one is concerned about health and about environmental effects. If tailings dams are allowed to dry out, and are not protected, wind is likely to generate a major dust problem, both a potential health and a potential environmental concern. Although the permeability of the very fine particles may be low, the very small particle size greatly enhances the surface area of the particles, and hence the contact area between the particles and any through flowing water: Especially when water flows through the particles at extremely slow rates, over prolonged periods of time (in the extreme, over centuries), it may interact sufficiently with some of the minerals in the particles to generate acidic outflows, as is the case in some districts where mining was conducted millennia ago. Such acidic releases frequently also are carriers of heavy metals dissolved in the liquid.

A third major source of solid mine wastes are spent heap leach dumps. Dump heap leaching is used primarily to recover gold and copper from low-grade ores, ores of which the metal content is insufficient to warrant considerably more expensive processing that requires crushing, usually grinding, and chemical or pyrometallurgical (high temperature) recovery of the metals. In its simplest form heap leach dumps are constructed by dumping run-of-mine ore into large heaps or dumps, on which a solution is sprayed that dissolves the metal of interest from the ore. Unfortunately all practically used leach solutions are toxic and environmentally damaging. The most commonly used solvents are sodium cyanide for gold and sulfuric acid for copper. Even though these are used in extremely diluted solutions, there remains a risk of toxic consequences unless controlled carefully. Once the leach dumps are spent, i.e., all recoverable metal values are removed, the dump material will be disposed of as waste. These materials constitute one of the larger sources of solid waste generated by many open pit gold and copper mines.

Underground mining usually generates only relatively small quantities of waste, at least compared to surface mining operations. One source of barren rock frequently generated in underground mining operations is the rock mined to develop access to ore bodies. This may involve shafts or ramps from the surface. It may involve underground mine development, e.g., drifting in barren rock toward the ore to be mined. It may involve the excavation of support facilities, e.g., underground equipment maintenance rooms, pumping stations, and crusher stations, i.e., the numerous underground space requirements for the efficient operation of a mine. Sometimes it is possible to leave some of this barren rock underground, e.g., in stopes that have been mined out previously as part of the ore production. Usually, this requires careful planning, and further complicates the sequencing of mining operations. Even so, the industry has clearly learned and understood that waste disposal on the surface involves costs, including considerable public perception costs, visual impact costs. Hence it seems probable that ever more effort will be devoted in the future to minimize the amount of solid waste that needs to be disposed of on the surface.

Quantities of Mining Solid Wastes

Quantities of solid waste are extremely large, voluminous. Consider a typical low-grade copper deposit that contains 0.5% copper (many deposits are being mined that contain significantly less, some that contain significantly more). Assuming a complete recovery, which is never the case, the production of 1 kg of copper requires processing 200 kg of such an ore. Hence 199 kg of the ore processed will remain as solid waste. Assuming a stripping ratio (i.e., ratio of barren overburden rock that needs to be removed in a surface mine in order to reach the mineral of economic value) of 2, a relatively low stripping ratio, 400 kg of overburden needs to be removed in order to access 200 kg of ore. In this case, the 1 kg of copper production results in generating 599 kg of waste.

Given the extreme diversity of the mining industry, it is exceedingly difficult to quantify the amount of waste it actually generates or has generated in the past, for reasons addressed briefly later in this section. In part, certainly, this may be due to the fact that until fairly recently waste management was not a priority for the producers nor for society at large. Only in the last few decades have the problems associated with solid mine waste disposal become fully recognized and appreciated, and, in particular, received extensive public and regulatory attention and scrutiny. Even though it can be assumed and expected that data gathering will improve in the future, for reasons explained in somewhat more detail in section on Future Directions, predicting how much solid waste will be generated by mining in the future also remains fraught with uncertainty.

In 1997, UNEP estimated that the solid waste annually generated by mining a group of selected major metals and industrial minerals amounted to about 2.7 billion tons, resulting from the mining of about 3.2 billion tons of ore [2]. However, this explicitly did not include any overburden mined; hence this almost certainly considerably understates the total amount of "waste" generated by these operations. It also did not include either coal or construction materials such as sand, gravel, dimension (building) stone, or crushed rock.

A reference document by the European Commission [3] estimates that approximately 118,000 kt of solid waste is generated annually by mining and quarrying in the EU-15, or about 20% of the solid waste produced in these countries. It deserves pointing that, as the reference explicitly recognizes, these "statistics on mining waste always bear a level of uncertainty," and, in all probability, this could be strengthened to a "considerable uncertainty." For example, the statistics cited, for 14 countries, are collected from different years, ranging over a 6-year period (1993–1999), because they simply are not available, for many countries, for every year, or even for many years. It needs to be recognized that Western Europe is a relatively minor mining waste producer, certainly compared to most large mining countries, and hence these numbers do not fully suggest the scale of the solid mine waste generated worldwide.

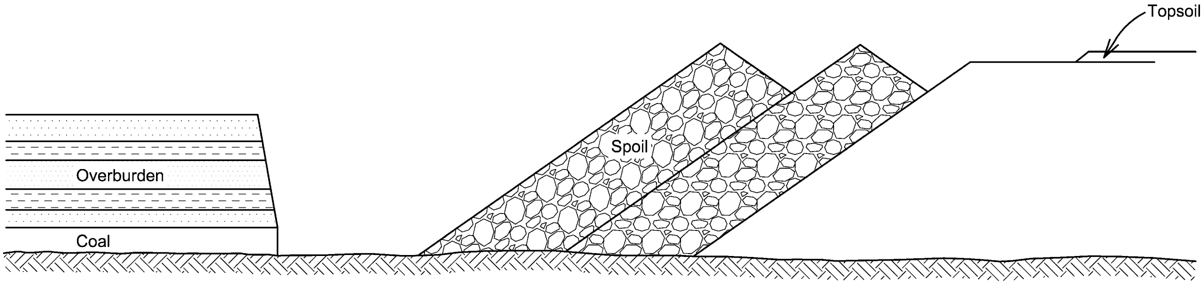
An exceptionally good explanation is given in references [4] and [5] as to why it is so difficult to determine how much solid mine waste exists, and how much is generated. Major data collection agencies come up with vastly different numbers (this holds true not only for solid mine waste, but for most other waste types as well, e.g., municipal solid waste, hazardous waste, demolition waste, etc.). Part of the uncertainty is due to the utter confusion and inconsistencies in waste definitions, e.g., between various governments, government agencies, data collection agencies and groups, etc. This uncertainty reflects rather well how recent it is that the waste problem has received due attention and the fact that people probably only are in the very early stages of dealing with it more or less adequately – given that the problem clearly is not yet fully or well defined. Although efforts are being made toward a more standardized and uniform definition of wastes, these will take time to implement and adopt. Moreover, significant differences exist in definitions and classifications between various governmental and regulatory bodies, and some of them are embodied deeply in legal frameworks, and hence likely to be very difficult to change. On top of that, in many countries and societies, especially poorer ones, waste management tends to remain a fairly low priority.

Storage/Disposal of Mining Solid Wastes

Storage of solid mine waste, i.e., temporary short-term management of solid mine waste, is rather unusual. Much more common is the permanent disposal, even though it may require maintenance. The type of disposal methods implemented will be closely intertwined with the mining method used, and will be influenced strongly by environmental and sometimes health-protection requirements.

In strip mining, a major surface coal mining method, long rectangular strips of coal are mined parallel to each other and directly adjacent to each other. Typically, the overburden from one strip is dumped into the opening left by the mining of the previous strip. Figure 1 illustrates schematically how the overburden is removed from above a coal seam and moved into the parallel pits left void after previous removal of the coal. Relatively shortly after the overburden spoil is dumped into the mined out strips, it is graded, leveled, and shortly thereafter topsoil is placed on the graded spoil. At this time revegetation can start. In these types of operations typically the vast majority of the overburden is replaced in mined out space. This method leaves separate disposal problems and challenges for the disposal of the overburden mined to access the first strip, and the fact that no overburden is available to refill the last strip. Particularly in areas with predominantly flat terrain, e.g., the Midwest of the USA, it has frequently been the practice to re-contour the spoil pile from the first strip into a hill, or hills. It is common practice to allow the last strip to fill with water, and to landscape it as desired, and, e.g., make it into a recreational lake, a reuse of the mined out area that is particularly attractive in areas where lakes are scarce.

The overburden piles, waste dumps, associated with open pit mining typically are left as barren rock piles, and are re-contoured in order to allow reclamation, i.e., revegetation. This usually will require that the slopes not be too steep, that topsoil be replaced and be used to cover the waste dumps, in order to provide a medium for plant growth. In recent decades there has been a growing interest in and movement toward backfilling of open pits, e.g., with overburden materials mined elsewhere in the pit, or sometimes in nearby pits. This somewhat controversial approach raises a serious question of resource conservation: Is it possible that, by this practice, potentially valuable ore resources may be covered up that will be considerably more difficult and expensive to mine at some point in the future? Mining history is replete with examples where further technological developments allowed mining of deposits that previously were abandoned as uneconomical. It is fairly routine, common, for historic mining districts to have gone through repeated mining cycles, sometimes with very lengthy dormant periods in between. But conversely there is no doubt but that pit backfilling reduces the visual impact of mining, and reduces the footprint of the impact left by mining, e.g., by greatly reducing the area occupied by solid waste disposal dumps.



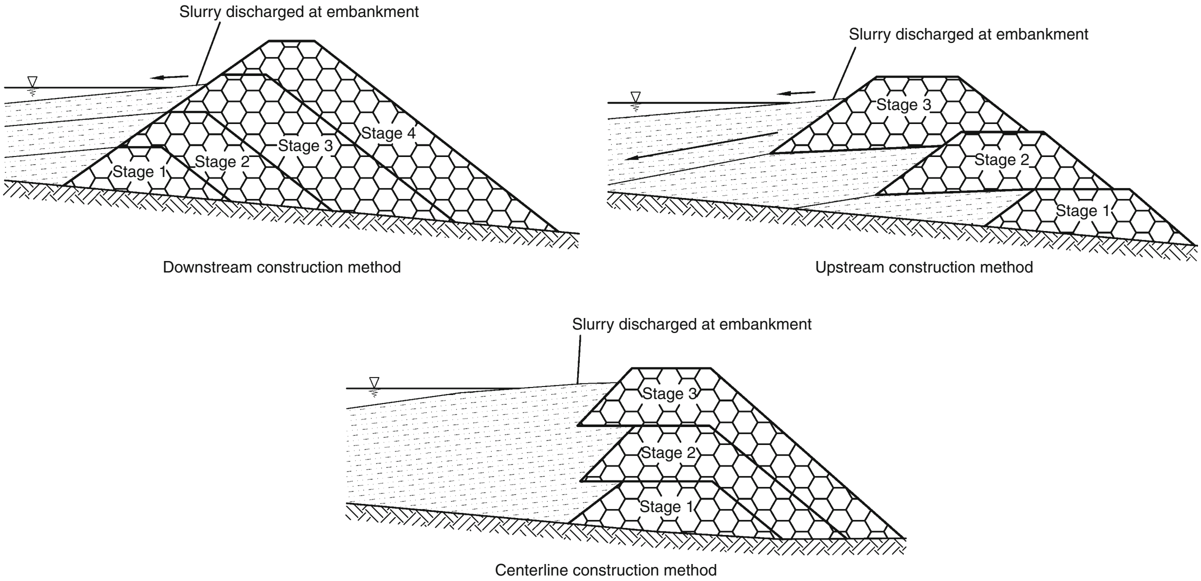
Mining Solid Wastes. Figure 1

Coal mined by strip mining, in which long parallel rectangular strips are mined, perpendicular to this vertical section through the pit. Overburden is stripped from above the coal (*left*) and dumped in the pits where the coal has been removed (*right*). The spoil piles are graded, a layer of topsoil is replaced on top of the graded spoil (*extreme right*), and reclamation can start, and can be performed concurrently with continuing, ongoing, mining (e.g., by seeding and planting)

Tailings disposal poses a particularly serious challenge. Tailings are very finely ground particles. Overwhelmingly they are transported as slurry, i.e., the particles are suspended in water, and pumped through pipes to the disposal dams or ponds. The resulting structures, some of the largest artificial dams in the world, pose major challenges to geotechnical engineers. By definition these structures, consisting of very fine particles, and, for some time after deposition, fully saturated, are intrinsically potentially highly unstable. Hence it is not surprising that these types of structures have resulted in multiple serious failures, including many truly catastrophic ones, with major loss of life, and with large detrimental environmental impacts. There is no doubt that these types of failures have been major contributors to the perception that mining is dangerous and damaging to the environment.

Figures 2 and 3 illustrate typical widely used methods of tailings dam construction. The basic difference between these methods is the sequence, the direction, in which the containment embankment is raised higher. For all three methods, the tailings are slurried to the disposal site, i.e., are carried suspended in water and pumped through pipes. The tailings are discharged a short distance behind the containment embankment. Typically the coarser particles will settle relatively close to the embankment, while the finer ones will travel to greater distances. The result is the formation of "beaches" in the disposal structure. It may be observed that in the second emplacement mode in Fig. 2, the upstream emplacement, subsequent embankments are built on top of these beaches. Hence the strength of the beaches has to be adequate to sustain the later additional loads imposed. In all cases the stability and impermeability of the containment structure are fundamental to assure adequate containment.

In order to minimize the risks associated with tailings dams and tailings ponds it is essential that such structures be planned, designed, constructed, and monitored with extreme care, by highly qualified professionals. There is no doubt that for many decades such structures were slapped together with extremely little if any care, with little understanding of the basic mechanics that determine their stability, or lack thereof, and of the high risk of failure of such structures. Over the last several decades major research efforts in many parts of the world have provided a greatly improved understanding of the failure modes involved, and of the design and construction requirements that need to be imposed in order to prevent future catastrophes. Unquestionably one of the serious challenges in this regard is that waste

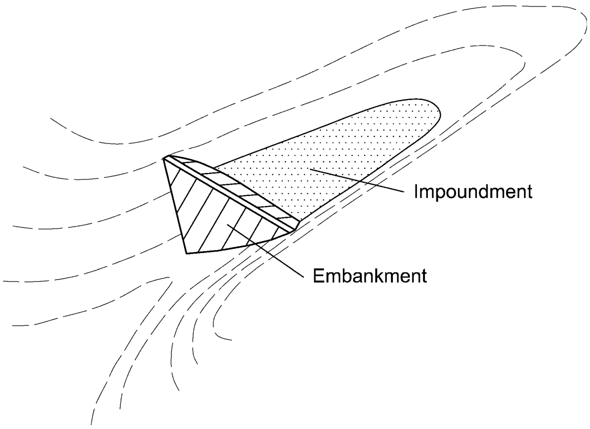

Slurry discharged at embankment
Stage 4
Stage 3
Stage 2
Stage 1
Slurry discharged at embankment
Stage 3
Stage 2
Stage 1
Slurry discharged at embankment
Stage 3
Stage 2
Stage 1

Downstream construction method

Upstream construction method

Centerline construction method

Mining Solid Wastes. Figure 2
Typical refuse facility (tailings dam) construction methods [6, Fig. 2, p. 3]



Mining Solid Wastes. Figure 3
Cross-valley tailings dam configuration [6, Fig. 1, p. 2]. A containment embankment is constructed across a valley. Tailings are pumped into the impoundment upstream of the embankment

disposal is a pure cost, and hence may not receive the necessary management attention to assure that it be done correctly. On the other hand, catastrophic failures of the types that have not been uncommon even in the last few decades carry a major penalty that can have a significant impact on the viability of a mining company. Aside from the ethical responsibility for mines to not endanger people or nature, the high visibility and impact of such failures should be an additional strong incentive to provide the necessary engineering to assure that such failures no longer will take place. One hopes. A more detailed and extensive argument and supporting discussion about the necessity and importance for the mining industry as a whole to reduce the number of catastrophic failures that draw great public attention because of their high visibility (and tailings dam failures have been among the most consequential and attention-grabbing disasters) has been given elsewhere [7].

Environmental and Health Concerns Related to Mining Solid Wastes

Depending on the waste characteristics and properties, a variety of environmental and health concerns may be associated with solid mine waste. To a major extent the concerns will be related to the chemical composition of the waste, sometimes to the physical characteristics, especially particle size.

Undoubtedly the major and most widespread, most common, and usually most serious environmental concern created by solid mine waste is the risk of acid drainage. If the mineralogy of the waste is such that some of the constituents of the waste are susceptible to interactions with water and oxygen that might result in the generation and release of acid, this will be of major concern to the public, environmental regulators, and the mine. Although in principle, conceptually, simple, e.g., a simple oxidation reaction of pyrite, the usual experience has been that the actual reactions and interactions taking place are far more complex than seems to be implied by such a simple formula. Particularly complicating dealing with this problem is that the reactions are affected by numerous variables, ranging from the obvious such as temperature and pressure to far from obvious and not yet fully understood synergistic effects between multiple minerals, varieties of water chemistries, and complicating biological and bacterial effects. Acid drainage has been the subject of major research efforts in virtually all major mining countries. Many successful remediation practices have been implemented, but problems still are encountered, and for specific situations further research may be needed.

Frequently closely related to acid mine drainage, associated with acid mine drainage, is the issue of release of heavy metals, some of which are highly toxic to humans, animals, fish, etc. Their immobilization and control will be a high priority objective of solid mine waste management.

In addition to chemical deterioration, physical effects on the environment may be of concern. A major issue in this regard is the presence of fines, especially materials less than 1 mm, in many solid mine wastes. One major concern about fines is that they can enter water streams or bodies, and result in sediment, reduced water clarity, and siltation, i.e., a variety of physical effects that can degrade surface water quality. In order to prevent such problems, it is essential that stringent sediment control and release practices be implemented at all mine waste handling facilities that are prone to sediment releases.

A somewhat related problem, also involving fines, is that of dust generation. Particularly for tailings dams and ponds, especially some years after they have ceased operations, i.e., after tailings deposition has been terminated, it is likely that the surface will dry out, certainly so in dry or semi-arid climates. Unless adequate preventive measures have been taken, it is rather probable that this might result in severe dust problems, especially in dry windy climates. If, moreover, the dust is contaminated with potentially toxic components, the problem obviously is compounded further. In sum, and as an example, uranium mill tailings in the dry and windy Southwest of the USA (e.g., New Mexico, Utah, Colorado), unless controlled carefully, could be causes of major environmental and health impacts, and over large areas.

Of particular concern, both with regard to environmental impact and with regard to public health and safety, is the question of the stability of tailings dams. Tailings dam failures have been the most catastrophic types of mine waste failures over the last century, both in regard to fatalities, public as well as mine personnel, and with regard to environmental impact. Major tailings dam failures have been associated with coal mine residue wastes (spoil piles), with copper mine tailings, with gold mine tailings, and with a variety of other minerals.

Intrinsically, tailings dams are potentially unstable. They consist of fine particle assemblies, extremely large assemblies, typically, during construction, emplacement, in a high degree of saturation. It is interesting to note in this context that some of the most catastrophic dam failures (civil construction dam failures) occurred at dams that were being or had been constructed as hydraulic fill structures, following a construction method essentially similar to the usual methods of tailings dam construction. While geotechnical engineers have learned a great deal from such failures, it often has been at an exceedingly steep price. Given the inherent potential instability of tailings dams, notably their high susceptibility to liquefaction, of particular concern is the risk that they might be subjected to dynamic earthquake loading. Liquefaction of saturated sandy soils has been a major cause of extensive damage induced by earthquakes on multiple occasions. Therefore it is essential that the dynamic stability of tailings dams be evaluated, certainly in areas with a relatively high earthquake risk.

It certainly can be argued that over the last century a great deal has been learned about the mechanics of tailings dam failures, and that it probably is correct that such structures can be planned, built, operated, and closed today without undue unacceptable risk. One critical factor remains, beyond any doubt: In order for a tailings dam to be safe, it needs to be constructed correctly. No matter how good the planning and design, deviations from the designed construction approaches during operations can easily result in structures that are far weaker, far less stable, than intended. Supervision, inspection, monitoring, observation during construction, by professionals fully and intimately familiar with the details of what makes a tailings dam stable or unstable may well be essential prerequisites to assuring fully satisfactory performance for such delicate structures. While this may seem obvious to outsiders, a self-evident truism, insiders will recognize that in the past, lack of qualified on-site engineering almost certainly has been a major contributing factor, if not a dominant cause of several catastrophic tailings dam failures, including some within the last few decades. It historically has not been perceived as essential, critical, that detailed structural aspects of such containment facilities must be handled correctly, with far less margin of error, uncertainty, than sometimes understood or accepted to be the case.

Use of Mining Solid Wastes

Compared to the overall volumes of waste generated, relatively little use is made of most of the solid waste generated by mining. In general, the volumes are so large, and the locations so remote, that not much use for solid mine waste is readily available.

One very important exception is the use of solid waste for underground mine backfill, although, again, relative to the total waste volumes generated by mining, this undoubtedly is a very minor fraction. Even so, for a number of situations this is an important use of solid mine waste.

Solid waste is backfilled in underground mines for a variety of reasons, but most commonly in order to improve ground control, i.e., to provide greater stability of underground openings. Waste can be backfilled for this purpose in a variety of ways. One common method is to dump relatively coarse rock particles, sized in the order of centimeters, into underground void spaces previously mined out. In order to strengthen the backfill, a small fraction (of the order of 5%) of cement or fly ash frequently is added to the backfill. The emplaced backfill applies confinement to the walls of the excavations, thereby greatly strengthening them, and reducing the risk of failures of the excavations. In the simplest of concepts, filling up the void space prevents the adjacent rock from bulking into the void space, thereby already greatly improving stability.

Fines, primarily tailings, may be backfilled in the form of a sand slurry, or as a viscous paste. When backfilled as a slurry, drainage of water will be an essential requirement in order to allow densification and strengthening of the backfill. Cement or fly ash here also may be added to give additional strength and stiffness. Backfilling of tailings by pumping them into mined out void space is a relatively simple and straightforward operation, but usually requires extensive preparatory construction of containment barriers that will assure that the tailings remain in place, in the intended locations, and that water can be removed, drained. Tailings emplaced in this fashion have similar geotechnical characteristics as tailings emplaced in dams, and hence raise similar stability concerns, notably with regard to dynamic (e.g., earthquake) loading.

Waste backfill has been used extensively in underground mines, especially in Europe, in order to reduce surface subsidence induced by mining. It was a common practice in coal mines in France, Belgium, and the UK, and still is used widely in Germany. It has been used in salt mines as well. One widely used method to backfill underground voids left by coal mining is to blow in fines with compressed air. Backfilling under these conditions may reduce surface subsidence by up to 50%, possibly more with a very dense (hence even more expensive) backfill.

A major benefit of returning solid mine wastes underground is that it reduces the visual impact of waste disposal on the surface. It reduces the space, e.g., land area, required on the surface for waste disposal. An environmental concern and risk about underground waste disposal is the potential for groundwater contamination. Hence, similarly to surface disposal, a comprehensive understanding of such risks is a prerequisite for decision making with respect to the acceptability of underground waste disposal.

Although there has been considerable research, and there have been efforts at using solid mine waste more widely, this remains the exception, rather than the rule. There have been applications for road construction, for construction fill, and similar applications, but these are not widespread. Unquestionably one of the major reasons is that the major cost for construction materials of this type, aggregates, is transportation. Rarely are large open pit mines in an area where there is great demand for construction materials.

Even for mine wastes that at one time were used, e.g., for road construction, or as fill for housing developments, this frequently is no longer an option, because of the recognition that at least some of these waste rocks contain constituents, e.g., heavy metals, that make their use highly questionable in applications where public exposure might result. It might be too difficult to demonstrate convincingly that no risk or negligible risk is involved, assuming it can be done.

Remediation of Mining Solid Wastes

Given the extremely wide range of materials being dealt with, it is not surprising that the remedial actions required in order to control any detrimental health and environmental effects associated with solid mine waste will be varied, complex, difficult, and expensive. While mining has been going on for millennia, only for a few decades has attention been paid to the need to protect people and the environment from the impacts of solid mine wastes. Considerable progress has been made over the last several decades, and a vast research literature is available on the subject, as well as a considerable data and fact base about a wide range of actual implementation practices of remedial actions that have been taken, although sometimes with mixed success. It seems reasonable to expect that the success rate should continue to improve.

As always when dealing with these types of problems, the first step is to try to minimize any source terms that contribute to potential health or environmental effects. The next step will be to implement control procedures to limit any impacts to an acceptable level.

A fundamental shift has taken place over the last few decades in mine planning and design, driven by the recognized need to minimize negative impacts from solid mine waste. Whereas a few decades ago mine design and planning was aimed primarily at maximizing profit (although historically other considerations sometimes were taken into account, e.g., maximizing resource recovery, or pursuing energy (partial?) independence, or pursuing self-sufficiency in strategic materials), today mining engineers are taught that mines need to be planned and designed, always, with closure, reclamation, remediation as major objectives, not simply as an afterthought, to be addressed eventually, if and when the need might arise.

This planning and design, taking into account the post-mining fate of the site, greatly affects numerous decisions to be made during the planning and design exercises. Most countries now require, as part of the granting of mine permits or licenses, applications that include environmental impact statements, in which considerable and credible detail needs to be provided about how the operations will be closed and terminated, what the post-mining use of the site will be, and how detrimental environmental impacts will be mitigated.

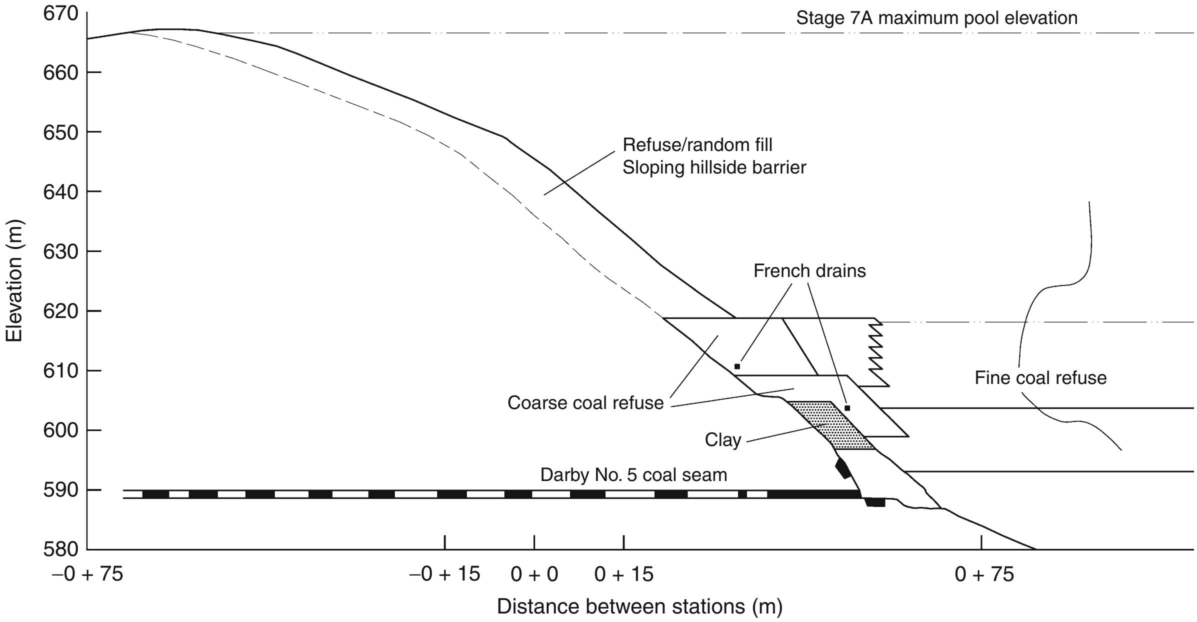
One important critical first step in such planning and design is site selection of waste emplacement. Whereas in the past the main if not only criterion in this regard was cost, convenience in locating waste disposal dumps, tailings dams, etc., as close as possible to the mining and milling operations, if only to minimize transportation costs, today such site selection needs to performed with great diligence.

For any waste that may incur a risk of water contamination, whether it be surface waters or groundwater, minimizing such risks will be a major site selection criterion. This will require consideration of surface and underground water flow patterns. It may require an environmentally acceptable rerouting of surface water courses. It may require isolating underground water flow, aquifers. Surface topography, permeability, and strength of the ground on which waste is to be emplaced will be investigated, tested, and will be significant factors influencing site selection.

One important aspect of environmental impact control at many mining operations will be waste, and in particular overburden removal and disposal management. For many ore deposits, mining operations, it is possible to consider the overburden as consisting of two primary components (this obviously is a rather considerable simplification): acid-generating waste and acid-consuming waste. An example of the first type might be a shale bed with a large fraction of pyrite. An example of the second type might be a relatively pure limestone bed. By judiciously planning and sequencing overburden mining operations, it might be possible to assure that all potentially acid-generating rock is encapsulated in, embedded by, acid-consuming rock. Such strategies contribute significantly to the reduction of acid generation. Although conceptually simple, following this exceedingly simplified description, in actuality the mine planner will most likely be dealing with a considerable range of rock types, with varying strength and weaknesses. Such mine planning, of necessity, needs to be based on a thorough mineralogical, geochemical, hydrological, etc., understanding of the potential for acid generation at the site. Environmental control planning of this type, unavoidably, fundamentally, requires a deep understanding of the driving factors, the fundamental scientific aspects of what creates environmental problems, in order to be able to mitigate against that potential.

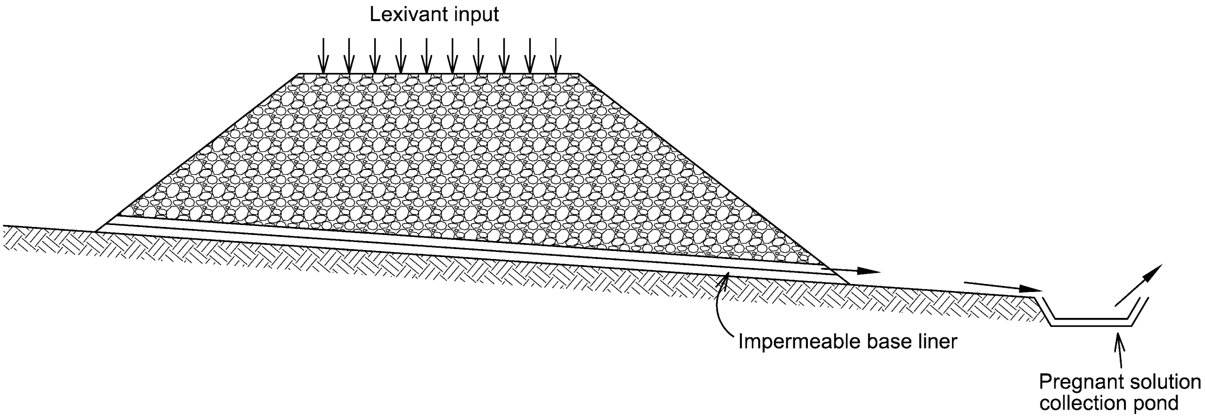
Many solid mine waste disposal facilities will be lined with relatively impermeable liners, and this certainly will be the case for the emplacement of all wastes that have the potential of generating acid drainage. Liners typically will consist of geomembranes, compacted clays, or a combination of both. Less common are asphalt or concrete liners. Installation of such containment structures is critically important for their performance, and hence needs to be conducted, supervised, and inspected with great care. Figure 4 illustrates the approach of using a clay barrier to prevent migration of slimes resulting from coal washing into an underground coal mine located directly underneath and adjacent to the refuse pile.

Closure of heap leach dumps will require first detoxification and next reclamation. Detoxification is performed by rinsing the heap leach dumps with chemicals that neutralize the solvents, react with the solvents, and convert them to innocuous residual chemicals, such as CO_2 and nitrogen gas (the quantities of CO_2 generated by these operations are trivial compared to those generated by, e.g., power production). One of the challenges for engineers closing heap leach dumps is to try to assure that all parts of the dumps have been rinsed and cleaned. It remains difficult, given the complexity of flow paths through such structures, to fully guarantee that no parts of the dump have been



Mining Solid Wastes. Figure 4

Cross section showing earthen liner to be constructed around a refuse impoundment to be installed near an underground mine in order to prevent the slimes, residue from washing the coal, from flowing into the mine (Modified from [6], Fig. 4, p. 7)



Mining Solid Wastes. Figure 5

Cross section through typical heap leach pad, as used for leaching gold and copper. Broken (blasted, sometimes crushed) rock is stacked in piles, in horizontal lifts of typically about 5 m thick, possibly up to a height of 50 m. Lexivant (leachate) is sprayed, sprinkled, or dripped on top. Pregnant solution is collected at the bottom. An impermeable pad is constructed on the foundation in order to minimize the risk of leachate escape to the groundwater

missed, that no pockets of cyanide or acid have been left behind. The problem is not particularly serious for cyanide, because it relatively quickly decomposes naturally in the presence of oxygen and water. For acid it is relatively straightforward to construct acid-consuming barriers, e.g., lime and limestone, and hence to assure that no acid can exit the site. Figure 5 illustrates a typical heap leach dump operation.

The crushed rock containing the metal to be dissolved (most likely gold, silver, or copper) is stacked loosely on an impermeable base liner. Leach solution is sprinkled, sprayed, or dripped on top of the dump, percolates through the dump, dissolves the metal(s) of interest, and is collected as the pregnant solution, i.e., the metal-bearing solution, in a nearby pond. From the pond it is pumped to a nearby plant where the metal of interest is recovered.

As for all waste containment structures, for heap leach dumps also the first step in minimizing the risk of inducing unacceptable releases will consist of selecting an appropriate site. Clearly attractive are sites where the natural base for the pads to be constructed, i.e., the soil or rock foundation, is impermeable and is sufficiently strong to withstand the load that will be exerted by the dump material in order to prevent, e.g., differential settlement across the base of the pad. Such differential settlement could induce cracking, shear or tensile failures, and thereby facilitate leachate flow through soil as well as through excessively deformed synthetic, i.e., geomembrane type liners. If an adequate site has been found, the next step will be the construction of the base-liner system, which will include an impermeable or very nearly impermeable barrier, above which a drainage system is installed that will allow draining the leachate and collecting it in a pond where the pregnant solution, i.e., the solution loaded with the valuable metal(s), will be gathered and from where it will be pumped to the metal recovery plant. All such leach pad systems installed will be provided with leak detection systems, designed and built in such a way that early warning is provided of any leaks occurring, in order to assure that timely remedial action can be taken.

Subsequent to detoxification of heap leach dumps, the further reclamation steps will be very similar to those of waste or overburden rock dumps. In all likelihood this will involve re-contouring, regrading, reshaping the dumps in order to facilitate revegetation, and in order to make them visually less obtrusive, e.g., by shaping into hill shapes very similar to those naturally occurring in the area. (In areas with flat terrain, it frequently is considered more attractive to maintain a hilly, "mountainous" configuration, and, e.g., provide the local communities with their own "mountain," most likely eventually to be used for recreational purposes.)

Once the desirable shape of the structure has been achieved, the next step will be to grow vegetation, and, e.g., to provide habitat for wildlife. The first step in this direction usually will be the emplacement of a layer of topsoil. In all operations of this type the topsoil, removed during the early stages of site preparation, will have been stored separately, for eventual reuse. Depending on the quality of the topsoil in semi-arid or desert climates, usually very poor and depending on local climatic conditions, the quality of the topsoil may, at the time when it is re-installed, be barely or not at all sufficient to allow plants to grow. Where this is the case, the soil will be amended as required, in order to provide an adequate growth medium. The usual preference today is to reclaim with native vegetation, either with seeding or planting, or a combination of both. It generally is considered desirable to establish a representative mix of plants appropriate for the area. This may require fertilization, and may require watering. In conjunction with the revegetation it is common to provide structures, e.g., an assembly of rock blocks with an arrangement known to be attractive to local wildlife. Wildlife biologists frequently, almost routinely, are employed by mining companies in order to assist with habitat recreation, optimized to accelerate the reestablishment of a diverse wildlife population.

Reclamation of quarries, mines used for the production of construction materials such as crushed rock, e.g., for road base and subbase, for Portland cement concrete or asphalt road concrete, is likely to be very different from that of coal or metal mines in remote isolated locations. In general, most quarries have no problem with chemical reclamation or contamination issues: Rocks that contain sufficient sulfides to be of concern from an acid generation point of view will almost never be acceptable for construction purposes. Hence, most deposits mined for construction material purposes have no associated geochemical contamination issues.

Because typically the major cost component of construction material crushed rock is transportation, many quarries, especially large ones, are started up relatively close to large urban areas. A classical historical pattern, repeated over and over again, is that over the life of the quarry the urban area continues to grow, and eventually totally encroaches on the quarry operation. Eventually the density of nearby population reaches the point where it becomes ever more difficult to operate the quarry, e.g., as a result of increasingly frequent complaints about noise and dust. The value of the land continues to grow, and reaches a magnitude where the land value exceeds, often by a considerable margin, the value of the remaining “ore” deposit, if any. In many locations around the world at this point the decision is made to convert the mine, the quarry, into a real estate development venture. If the quarry was mined sufficiently deep (and it does not have to be very deep at all), it may be possible to convert the quarry into a lake, e.g., for recreational purposes, or as focal point of urban development, sometimes residential, sometimes commercial, sometimes a combination of both.

Undoubtedly one of the major challenges facing many societies is that of dealing with legacy sites, polluted sites remaining from mining operations in the distant past, that have left major massive contamination problems at a number of locations, in many countries with longstanding mining histories, especially metal mining, but certainly also surface coal, lignite, brown coal, or soft coal operations. In most of these situations remedial actions will require, or have required, massive cleanup efforts, on a vast scale.

Future Directions

As repeatedly stated, the generation of large volumes of solid waste is an unavoidable side effect of mining. Hence the first step in estimating the future of solid mine wastes is to make a brief estimate of the future of mining. For a number of reasons it is to be expected that society at large will need more minerals, and hence that mineral production will expand. The most obvious reason is that a large fraction of the world population continues to live in subsistence poverty levels, and that efforts will continue to improve the standard of living of a larger fraction of the world population. This will require more mineral production. Concurrent with this trend is the growing standard of living for a rapidly growing middle class, worldwide, notably in the largest players, China and India. As the standard of living continues to improve for these hundreds of millions of people, they will need more steel (e.g., for cars and buildings), more construction materials (e.g., for roads, buildings, and houses), more copper (e.g., to wire their houses and cars, and to distribute electrical power), more precious metals (e.g., platinum for car catalytic converters and jewelry, more gold and silver for electronics but especially as jewelry), and more energy (e.g., coal and uranium). In sum, all indications and predictions are that mine production will increase.

Mine production, in volume of materials mined, also will have to increase. For metals this is fairly obvious, as the grades of ore deposits that are being mined continue to decrease, and hence more solid waste needs to be produced per unit of metal recovered. Although for decades professionals have argued that more mining is likely to move underground, the reverse actually has happened: open pit and surface strip mines have continued to grow in size, most notably by going ever deeper, as has been made feasible by ever larger, more powerful, and more efficient surface mining equipment.

The picture is somewhat less clear for coal, one of the major sources of solid mine waste. In the USA as well as in other parts of the world, the future of coal is less obviously one of steady growth than it seemed 10 years ago. The vast expansion of natural gas discoveries and production is putting major price competition in the path of coal growth. Simultaneously, concurrently, in parallel, major efforts are being made to reduce greenhouse gas emissions, and it seems rather likely that this eventually will result in reduced use, or at least in a reduced growth of the use of coal for electrical power generation. Nevertheless, for the foreseeable future, it appears likely that coal will continue to be a major source of electrical power.

All indications are that the revival of nuclear power, most obvious in China and India, will continue, with a concomitant growing demand for uranium. Again, the scale to which this trend may proceed remains difficult to predict, but nevertheless it seems highly probable that uranium mining and milling, with their unique and serious environmental control requirements, will continue to grow.

In light of this high probability of increased solid mine waste production, it is apparent that the need for improvements in the controls on environmental impacts of such waste will increase as well. Major advances have been made over the last few decades in improving the understanding of the causes and mechanics of environmental degradation caused by or associated with solid mine disposal. Mine planning and design have been fundamentally altered, with as a major objective making mining practices more environmentally acceptable. It seems certain that this trend will continue, probably intensify, and become more widespread, i.e., become more widely adapted.

Although the understanding of the fundamental causes of environmental degradation, e.g., the chemistry of acid drain formation, the physics of particle releases and consequent sedimentation, heavy metal transport, has improved greatly, it also has become clear that many of the fundamentals remain understood only partially, and are in need of further study. Research on this subject continues in virtually all major mining countries, and undoubtedly will result in further improvement in the understanding of the problems, and in the development of technologies for dealing with them.

As experience continues to be gained with control measures, e.g., the use of barriers aimed at containing potential contaminant plumes, such technologies will continue to improve, continue to result in approaches that are more effective, and can be implemented at lower cost. Of particular importance, given that major concerns about these issues have developed only over the last few decades is that as experience is gained over decades, a much better understanding is being developed of the long-term impact of potential health and environmental effects from solid mine waste, about the effectiveness, or lack thereof, of various prevention and control strategies, and of the need for future and further improvements in methods and strategies for dealing with such challenges.

Numerous solid waste disposal facilities have been installed over the last few decades with engineered containment structures and with engineered containment approaches. Many of these facilities have been equipped with monitoring instrumentation. A variety of monitoring methods have been implemented, such as leak detection systems in or below liners, water quality sampling and monitoring wells near dumps, and air sample collection instrumentation. It is certain that over the next few decades a vast experiential database will be collected, an extensive set of factual data at actual disposal operations.

In parallel, ongoing major research efforts aimed at elucidating the fundamental causes of potential environmental impact problems caused by solid mine wastes are seen. This includes detailed laboratory studies of the complex interactions that take place within and on the surface of waste dumps, tailings dams, and spent heap leach pads. Combined with laboratory studies are increasingly sophisticated and detailed numerical modeling studies, computer simulations of the reactions, and water flows that take place in, on, and near such structures. The extensive and intensive full-scale experiments that are being conducted, in many different places, under a wide range of climatic conditions, dealing with a wide variety of solid wastes, will produce comprehensive databases that will allow calibrating and assessing numerical codes and the predictive validity of a variety of laboratory tests and simulations.

It seems virtually certain that the combined and synergistic results of all these efforts will continue to improve the understanding of what takes place inside solid mine waste disposal facilities, and hence will lead to improved approaches to reduce and minimize the detrimental consequences of such wastes, unavoidably generated by mining. The commitment from many major mining companies and from numerous research and regulatory agencies toward such a goal is firm and unequivocal.

Why is all this waste continued to be produced? Is there no alternative? Are there no mining methods that allow recovery without the production of all these massive amounts of waste?

On occasion, although mostly in the somewhat distant past, dreamers, or innovative thinkers have conceptualized mining methods, e.g., in situ mining, that bypass the need for the removal of the vast amount of overburden or barren rock that are typically used in mining. In fact, for a few minerals this is a common mining method. Minerals most commonly mined by in situ mining methods include sulfur and rock salt.

Sulfur is routinely mined by drilling boreholes into the sulfur deposits, injecting hot steam, under high pressure and temperature, thereby melting the sulfur, and then pumping the molten sulfur to the surface. Rock salt frequently is mined by drilling holes into the salt deposit to be mined, injecting water through the hole, allowing salt to dissolve in the water, and then pumping the salt-laden brine to the surface, where the water is evaporated, and the salt crystallized and recovered. These highly unusual mining practices are possible for these specific minerals because of some unique characteristic (i.e., low melting point, ready solubility). Rarely are such methods feasible for any other materials.

In situ leaching, dissolving metals of interest and value while leaving them in the ground, by accessing ore deposits with boreholes, has been of interest for decades. It is used on an extensive scale for uranium, for which it is a major mining method. It has been practiced on a few occasions for copper, although most commonly in association with conventional mining, e.g., in halos of low-grade ore near mined out richer deposits. It has been discussed for gold. A major concern for in situ mining, especially when considering the use of solvents, is the potential for environmental impact, notably the risk of groundwater contamination. While conceptually feasible, the installation of impermeable barriers around a site to be leached is not easy, certainly is expensive, and confidently assuring that they truly are and will remain impermeable would be a challenge.

In situ coal gasification has been demonstrated to be feasible in several countries. The method essentially consists of starting a controlled fire in a coal seam, and limiting the oxygen available for combustion to assure that useful gases can be recovered, especially methane and carbon monoxide. While technically feasible, the economics clearly have not been convincing for decades, and in this day and age it is likely that far more careful studies would be required of the environmental impacts of such methods.

The technical feasibility of in situ oil shale retorting has been demonstrated convincingly. By this method heavy viscous oil can be recovered from tight impermeable shale formations by inducing a controlled fire, raising the temperature, reducing the viscosity of the oil, increasing the rock permeability, thus allowing oil to flow, and to be recovered. As implemented so far, the method requires developing access along the top and along the bottom of the oil shale layer. This generates some solid waste, but certainly far less than if the shale were mined completely, and the oil recovery implemented on the surface, in surface retorts. Presumably when oil reaches the right price level, this method might be pursued as one option to generate energy, with less solid waste generation.

The most obvious and direct approach to reduce the need for mining, and the associated solid waste generation, is to further enhance and promote recycling. Recycling of metals has been practiced for many centuries, but has become more industrial based in recent decades, and has become driven more strongly by environmental concerns even more recently. Recycling of aluminum, long common and successful, historically has been pursued primarily because of the extremely favorable economics of aluminum recycling: large energy savings for recycling aluminum, as compared to production from raw ore, make it economically very attractive. To a much lesser degree, but nevertheless similarly driven, is the recycling of copper and steel. Recycling of silver and gold from electronic wastes is a well-established practice, as is the recycling of platinum and palladium from catalytic converters. In sum, metal recycling promises to continue to reduce the need for the production of virgin metal, and associated mine solid waste production. To date, it may have slowed the growth of new metal mining, but growth certainly still continues. Considering the large number of people who still live at a metal consumption level far below that of an adequate standard of living, it seems most likely that this trend will continue: recycling can substitute for some fraction of metal demand, but probably a relatively small fraction only, or at least only a fraction that does not yet make it possible to reduce mining new raw materials.

In a similar vein, one approach to reduce the need for energy mining, e.g., coal, uranium, and tar sands, is to strengthen energy conservation efforts. In parallel with the aggressive pursuit of renewable energy, this is likely to result in a reduction in the growth rate of the production of conventional fuels, and associated solid mine waste generation. To date, once again given the worldwide growth in energy demand, it appears that this will result in a reduced growth rate, not in actual reduced production.

Although an argument can be made that recycling of construction materials has been practiced since historical times, in many societies people have dismantled ancient structures, ruins, to reuse building stone, industrial scale recycling of concrete and asphalt is fairly new. Today it is being pursued quite aggressively in many societies. As the impetus for doing so strengthens, and, as a result, as the equipment and the methods for such recycling keep on being improved, it is likely that recycling of such basic construction materials as crushed rock, sand, and gravel, i.e., aggregates, will expand, and, correspondingly, will result in reduced demand for newly mined materials. This may be a significant contributor to reducing the demand for new quarries, in those societies where recycling of these materials can provide a significant fraction of the demand for such materials. It will have far less impact, if any, in societies where basic infrastructure still needs to be built.

Conceptually, one might conceive of a society less material intensive. And to a significant extent modern societies have achieved this. Modern electronics deliver far more than their far heavier precursors. Far less metal is needed for modern devices of many kinds than used to be the case. To date this is not reflected in a reduced overall demand for metal, energy, or construction rock. While gadgets have become lighter, they also have become more numerous. Houses have become larger, requiring more concrete, copper, and glass. Cars have become larger, at least in some societies, requiring more steel and copper. Unless societal preferences change rather drastically, a move toward a less materials-intensive culture does not appear imminent. As of now, it seems far more likely that the demand for minerals will continue to grow, but it seems equally likely that environmental control requirements will continue to intensify on the producers of the minerals societies desire and need.

According to its 2001 Environment Outlook [8, p. 238] the OECD expects that in OECD countries "...waste from the primary sectors, such as ... mining and quarrying, is expected to grow at a slower rate" (i.e., slower than other wastes, such as municipal waste and manufacturing waste). Even in advanced societies solid mine waste is still expected to grow, not decline. In an updated version of its projections [9, p. 239] OECD estimates that by 2020 metal ore extraction will increase from 5.8 billion tons in 2000 to more than 11 billion tons in 2020. On a per capita base, OECD expects that the solids extraction in OECD countries will reach 22 tons per person by 2020, and 9 tons in the BRIICS countries (Brazil, Russia, India, Indonesia, China, and South Africa). This, combined with the expected population growth in most of these countries indicates why it is virtually certain that the scale of the solid mine waste problem is likely to increase substantially over the next decade, and most likely for several decades beyond that, at least. OECD countries produce about 30% of metal ores, BRIICS countries nearly 40%, and the rest of the world slightly over 30%. Hence the impact of solid waste generated by mining is widespread.

It seems certain and obvious that societies will continue to become less and less tolerant for the types of environmental impacts that mining used to have. In order to successfully provide the minerals needed for people, it is essential that the mineral producers develop mining methods that allow the management of solid mining wastes without unacceptable detrimental consequences.

Acknowledgment

The authors would like to thank Mr. Kıvanç Okalp, Coordinator of Projects, Beray Engineering Consultancy Ltd. for his assistance in drafting the illustrations in this chapter.

Bibliography

Primary Literature

1. Nazzaro RM (2008) Hardrock mining-information on abandoned mines and value and coverage of financial assurances on BLM land, United States Government Accountability Office, Testimony GAO-08-574 T (United States Government Accountability Office, Washington, DC)
2. UNEP (1997) Mining – facts and figures. Ind Environ 20:4–9
3. European Commission (2009) Reference document on best available techniques for management of tailings and waste-rock in mining activities. http://eippcb.jrc.ec.europa.eu. Accessed 8 April 2010
4. Twardowska I, Allen HE (2004) Solid waste origins: sources, trends, quality, quantity, Chapter I.2. In: Twardowska I, Allen HE, Kettrup AAF, Lacy WJ (eds) Solid waste: assessment, monitoring and remediation. Amsterdam, Elsevier, pp 33–88
5. Szczepańska J, Twardowska I (2004) Mining waste, Chapter III.6. In: Twardowska I, Allen HE, Kettrup AAF, Lacy WJ (eds) Solid waste: assessment, monitoring and remediation. Amsterdam, Elsevier, pp 319–385
6. Michalek SJ, Gardner GH, Wu, KK (n.d.) Accidental releases of slurry and water from coal impoundments through abandoned underground coal mines (Mine Safety and Health Administration. Pittsburgh Safety and Health Technology Center, Pittsburgh, PA. www.msha.gov/S&HINFO/ TECHRPT/MINEWASTE/ASDSO2.pdf). Accessed 6 April 2010
7. Daemen JJK (2009) Geotechnical/geomechanical mine failures. International workshop on winning strategies to

revitalize the mineral sector, August 8–9 2009. Department of Mining Engineering, National Institute of Technology Karnataka, Surathkal, India, Proceedings pp 111–125

8. OECD (2001) OECD Environmental outlook environment (Organisation for Economic Co-Operation and Development, Paris, France)
9. OECD (2008) OECD Environmental outlook to 2030 (Organisation for Economic Co-Operation and Development, Paris, France)

Books and Reviews

Agioutantis Z, Komnitsas K (eds) (2006) In: Proceedings of the 2nd international conference on advances in mineral resources management and environmental geotechnology (AMIREG 2006), 25–27 September 2006. Chania, Crete, Greece, Heliotopos conferences, Athens

Aplin CL (1973) Tailings disposal today. In: Proceedings international tailings symposium (1st, 1972, Tucson, AZ). Miller Freeman Publications, San Francisco

Argall Jr GO (1979) Tailings disposal today. In: Proceedings of the second international tailings symposium, vol 2, May 1978. Denver, CO, Miller Freeman Publications, San Francisco

Aston RL (1999) The legal, engineering, environmental and social perspectives of surface mining law and reclamation by landfilling: getting maximum yield from surface mines. Imperial College Press, London

Azcue JM (ed) (1999) Environmental impacts of mining activities: emphasis on mitigation and remedial measures. Springer, Berlin/Heidelberg/New York

Beedlow PA, Gee GW, Cline JF, Walters WH, Freeman HD (1985) Determination of compliance with criteria for final tailings disposal site reclamation, technical report NUREG/CR-4076, PNL-5324. Prepared by Pacific Northwest labs, Richland, WA (USA) for the US Nuclear Regulatory Commission, Washington, DC

Bell FG (1998) Environmental geology: principles and practice. Blackwell Science, Oxford

Bennett RD, Horz RC, Kimbrell AF (1991) Recommendations to the NRC for soil cover systems over uranium mill tailings and low-level radioactive wastes, NUREG/CR-5432. US Nuclear Regulatory Commission, Washington, DC

Billings Land Reclamation Symposium (2000). In: Proceedings 2000 Billings land reclamation symposium: March 20–24, 2000, Billings, MT. Reclamation Research Center, Water Center, Montana State University, Bozeman

Bouazza A, Kodikara J, Parker R (eds) (1997) Environmental geotechnics. In: Proceedings of the 1st Australia-New Zealand conference on environmental geotechnics-geoenvironment 97, Melbourne, VIC, Australia, 26–28 November 1997. AA Balkema, Rotterdam/Brookfield

Broughton SE (1992) Documentation and evaluation of mine dump failures for mines in British Columbia, mine rock and overburden piles. Review and evaluation of failures, interim report, British Columbia mine waste rock pile research committee. Ministry of Energy, Mines and Petroleum Resources, British Columbia

Cabri LJ, Vaughan DJ (1998) Modern approaches to ore and environmental mineralogy, short course handbook, vol 27. Mineralogical Association of Canada, Ottawa

Chalkley ME, Couard BR, Lakshmanan VI, Wheeland KG (eds) (1989) Tailings and effluent management. In: Proceedings, The metallurgical society of the Canadian Institute of Mining and Metallurgy, vol 14. Pergamon Press, New York

Charles River Associates Incorporated (1985) Estimated costs to the US mining industry for management of hazardous solid wastes. Prepared for US Environmental Protection Agency, Division of Solid Waste Management, Washington, DC. Charles River Associates Incorporated, Boston, MA

Chaudhuri AB (1992) Mine environment and management (an Indian scenario). Ashish Publishing House, New Delhi

Coates DF, Yu YS (eds) (1977) Pit slope manual, Chapter 9-Waste embankments, CANMET (Canadian Centre for Mineral and Energy Technology, formerly Mines Branch, Energy Mines and Resources Canada, CANMET Report 77-1. Ottawa, Canada

Colorado State University Geotechnical Engineering Program (1980) Uranium tailings management. In: Proceedings of the third symposium, November 24–25, 1980, organized by the Geotechnical Engineering Program. Civil Engineering Department, Colorado State University. Colorado State University, Fort Collins, CO

Colorado State University Geotechnical Engineering Program (2001) Tailings and mine waste'01. In: Proceedings of the eight international conference on tailings and mine waste, Fort Collins, CO, 16–19 January 2001. AA Balkema, Rotterdam

Colorado State University Geotechnical Engineering Program (2002) Tailings and mine waste'02. In: Proceedings of the ninth international conference on tailings and mine waste, Fort Collins, CO, 27–30 January 2002. AA Balkema, Rotterdam

Commission of inquiry appointed by his excellency the President of the Republic of Zambia Dr. Kenneth David Kaunda (1971) The Mufulira mine disaster. Republic of Zambia, Lusaka

Commission of the European Communities (2003) Proposal for a directive of the European Parliament and of the Council on the management of waste from the extractive industries, 2003/0107(COD), COM(2003) 319 final. Commission of the European Communities, Brussels

Commission of the European Communities (2006) Directive (2006/21/EC) of the European Parliament and of the Council on the management of waste from the extractive industries, EN, L 102/15-33, 11.04.2006. Commission of the European Communities, Brussels

Committee on Embankment Dams and Slopes (1979) Current geotechnical practice in mine waste disposal. American Society of Civil Engineers, New York

Daniel DE (ed) (1993) Geotechnical practice for waste disposal. Chapman & Hall, London

D'Appolonia Consulting Engineers (1985) Engineering and design manual: Coal refuse disposal facilities. US Department of the Interior, Mining and Safety Administration, Washington, DC

Davies WE, Bailey JF, Kelly DB (1972) West Virginia's Buffalo Creek flood: a study of the hydrology and engineering geology. Geological Survey Circular 667, US Geological Survey, Washington, DC

DeGraff JV (ed) (2007) Understanding and responding to hazardous substances at mine sites in the western United States, reviews in engineering geology XVII. The Geological Society of America, Boulder

Denham DH, Barnes MG, Rathbun LA, Young JA (1985) Monitoring methods for determining compliance with decommissioning cleanup criteria at uranium recovery sites, Technical report NUREG/CR-4118, PNL-5361. Prepared by Pacific Northwest labs, Richmond, WA for the US Nuclear Regulatory Commission, Washington, DC

Dhar BB (ed) (1990) Environmental management of mining operations. Ashish Publishing House, New Delhi

Dhar BB (2000) Mining & environment. APH Publishing Corporation, New Delhi

DHI Water-Environment-Health (2007) Classification of mining waste facilities, European Commission DG Environment Final Report No. 07010401/2006/443229/MAR/G4. Prepared by DHI Environment Health in cooperation with SGI, Swedish Geotechnical Institute and AGH University of Science and Technology, Krakow. http://europa.eu/legislation_summaries/environment/waste_management/128134_en.htm. Accessed 6 April 2010

Doyle FM (ed) (1990) Mining and mineral processing wastes. In: Proceedings of the western regional symposium on mining & mineral processing wastes, Berkeley, California, May 30–June 1, 1990. Society for Mining, Metallurgy, and Exploration, Littleton, CO

Down CD, Stocks J (1976) The environmental impact of large stone quarries and open-pit non-ferrous metal mines in Britain. Department of the Environment and Transport Research Report 21, South Ruislip, United Kingdom

Down CG, Stocks J (1977) Environmental impact of mining. A Halsted press book. Wiley, New York

Environment Protection Agency (1995) Tailings containment. Australian Federal Environment Department, Barton Act, Australia

Erikson KT (1976) Everything in its path-destruction of community in the Buffalo Creek flood. Simon and Schuster, New York

Fourie A (2008) Rock dumps 2008. In: Proceedings of the first international seminar on the management of rock dumps, Stockpiles and heap leach pads, 5–6 March, 2008, Perth, Australia. Australian Centre for Geomechanics, Nedlands, Western Australia

Fourie A, Tibbet M (eds) (2009) Mine closure 2009. In: Proceedings of the fourth international conference on mine closure, Perth, Australia, 9–11 September, 2009. Australian Centre for Geomechanics, Nedlands, Western Australia

Fung R (ed) (1981) Surface coal mining technology: engineering and environmental aspects, energy technology review no. 71, Pollution technology review no. 83. Noyes Data Corp, Park Ridge, NJ

Gadsby JW (ed) (1990) Acid mine drainage: designing for closure. BiTech, Vancouver

Goin P, Raymond CE (2004) Changing mines in America, Center for American Places, Santa Fe, NM. Distributed by University of Chicago Press, University of Chicago, Chicago, IL

Golder Associates (1994) Mined rock and overburden piles. Consequence assessment for mine waste dump failures, Interim report, British Columbia Mine Waste Rock Pile Research Committee. British Columbia Ministry of Energy, Mines and Petroleum Resources, BC

Golder associates (1995) Mined rock and overburden piles. Runout characteristics of debris from dump failures in mountainous terrain, stage 2: Analysis, modelling and prediction, Interim report, British Columbia Mine Waste Rock Pile Research Committee, British Columbia Ministry of Energy, Mines and Petroleum Resources, BC

Hudson-Edwards K, Savage K, Jamieson H, Taylor K, Martin RF (eds) (2009) Minerals in contaminated environments: Characterization, stability, impact, The Canadian Mineralogist, Thematic Issue, Vol. 47, Part 3. Mineralogical Association of Canada, Québec

Hutchison IPG, Ellison, RD (eds) (1992) Mine waste management: a resource for mining industry professionals, regulators, and consulting engineers. Sponsored by California Mining Association, Lewis Publishers, Chelsea, MI

Hustrulid WA, McCarter MK, van Zyl DJA (eds) (2000) Slope stability in surface mining, section 3. In: Tailings and heap leaching. Society of Mining, Metallurgy, and Exploration, Littleton, CO, pp 265–362

Hustrulid WA, McCarter MK, van Zyl DJA (eds) (2000) Stability of waste rock embankments, Section 4. In: Tailings and heap leaching, Society of Mining, Metallurgy, and Exploration, Littleton, CO, pp 363–438

IAEA (2002) Monitoring and surveillance of residues from the mining and milling of uranium and thorium. IAEA (International Atomic Energy Agency), Vienna

IAEA (2004) The long term stabilization of uranium tailings: final report of a coordinated research project 2000–2004. IAEA (International Atomic Energy Agency), Vienna

IAEA (2010) Best practice in environmental management of uranium mining, IAEA Nuclear Energy Series No. NF-T-1.2, IAEA (International Atomic Energy Agency), Vienna

ICOLD (1989) Tailings dam safety: guidelines. International Commission on Large Dams, Paris

ICOLD (1994) Tailings dams design of drainage, review and recommendations. International Commission on Large Dams, Paris

ICOLD (1995) Tailings dams: transport placement and decantation. International Commission on Large Dams, Paris

ICOLD (1995) Tailings dams and seismicity: review and recommendations. International Commission on Large Dams, Paris

ICOLD (1996) A guide to tailings dams and impoundments: design, construction, use and rehabilitation, bulletin 106. International Commission on Large Dams, Paris

ICOLD (1996) Monitoring of tailings dams: review and recommendations. International Commission on Large Dams, Paris

ICOLD (2001) Tailings dams risk of dangerous occurrences: lessons learnt from practical experiences. International Commission on Large Dams, Paris

International Conference on Tailings & Mine Waste (2001) Tailings and mine waste'01. In: Proceedings of the eight international conference on tailings and mine waste'01, Fort Collins, CO, USA, 18–19 January 2001. AA Balkema, Rotterdam

Jacobs Engineering Group, Inc. (1994) Remedial action plan and site design for stabilization of the inactive uranium mill tailings sites at Slick Rock, Colorado, remedial action selection report, preliminary final, DOE/AL/62350-21PF. US Department of Energy, UMTRA Project Office, Albuquerque, New Mexico. www.osti.gov/bridge/servlets/purl/10135330-8eDj6i/.../10135330.pdf

Jambor JL, Blowes DW (eds) (1994) The environmental geochemistry of sulfide mine-wastes, Mineralogical Association of Canada short course series vol 22. Mineralogical Association of Canada, Neapan, ON

Jambor JL, Blowes DW, Ritchie AIM (2003) Environmental aspects of mine wastes, short course series, vol 31. Mineralogical Association of Canada, Ottawa

Koerner RM, Daniel DE (1997) Final covers for solid waste landfills and abandoned dumps. ASCE Press, Reston, VA and Thomas Telford, London

Kreft-Burman K, Saarela J, Anderson R (2005) Tailings management facilities legislation, authorisation, management, monitoring and inspection practices, TAILSAFE, http://www.tao;safe.com. Finnish Environment Institute (SYKE), Helsinki

Lindsey CG, Long LW, Begej CW (1982) Long-term survivability of riprap for armoring uranium mill tailings and covers: a literature review, NUREG/CR-2642. Division of Health, Siting and Waste Management, Office of Nuclear Regulatory Research, US Nuclear Regulatory Commission, Washington, DC

Lottermoser BG (2007) Mine wastes: characterization, treatment and environmental impacts, 2nd edn. Springer, Berlin

Lusher JH (2003) Standard review plan for the review of a reclamation plan for mill tailings sites under Title II of the Uranium Mill Tailings Radiation Control Act of 1978, Final report. NUREG-1620, Rev. 1, Office of Nuclear Material Safety and Safeguards, US Nuclear Regulatory Commission, Washington, DC. http://www.nrc.gov/reading-rm/doc-collections/nuregs/staff/sr1620/sr1620r1.pdf

Marcus JJ (ed) (1997) Mining environmental handbook: effects of mining on the environment and American environmental controls on mining. Imperial College Press, London

McCarter MK (ed) (1985) Design of non-impounding mine waste dumps. Society of Mining Engineers of the American Institute of Mining, Metallurgical, and Petroleum Engineers, New York

Mining Association of Canada (1998) A guide to the management of tailings facilities. The Mining Association of Canada, Ottawa, ON. www.mining.ca/english/ publications/tailingsguide.pdf

Mining Journal Research Services (1996) Environmental and safety incidents concerning tailings dams at mines: results of a survey for the years 1980–1996. United Nations Environment Programme, Industry and Environment, Paris

Montana Department of Natural Resources and Conservation Engineering Bureau (1977) Mine drainage control from metal mines in a subalpine environment-A feasibility study, EPA-600/2-77-224, Environmental Protection Technology Series. US Environmental Protection Agency, Industrial Environmental Research Laboratory, Office of Research and Development, Cincinnati, OH

MSHA (1979) Design guidelines for coal refuse piles and water, sediment, or slurry impoundments and impounding structures, Informational report 1109. US Department of Labor, Mine Safety and Health Administration, Arlington, VA

MSHA (2007) MSHA coal mine impoundment inspection and plan review handbook. US Department of Labor, Mine Safety and Health Administration, Arlington, VA. www.arblast.osmre.gov/.../MSHA%20PH07CoalImpoundmentinspection.pdf

Murray DR (1977) Pit slope manual, supplement 10-1-Reclamation by vegetation: vol 1-Mine waste description and case histories; CANMET (Canadian Centre for Mineral and Energy Technology, formerly Mines Branch, Energy Mines and Resources, Canada, CANMET report 77-31. Ottawa

Murray DR (1977) Pit slope manual, supplement 10-1-Reclamation by vegetation: vol 2-Mine waste inventory by satellite imagery; CANMET (Canadian Centre for Mineral and Energy Technology, formerly Mines Branch, Energy Mines and Resources, Canada, CANMET report 77-58. Ottawa

National Coal Board (1972) Review of research on spoil pile tip materials, Lab. Reference No. S/7307. Wimpey Laboratories, Hayes, Middlesex

Nash JT (2002) Flood deposits of transported mill tailings in Nevada and Utah: evidence for tailings dam failures and implications for risk assessment. US Geological Survey open-file report 02-431. US Department of the Interior, US Geological Survey, Denver, CO

Nash JT (2003) Historic mills and mill tailings as potential sources of contamination in and near the Humboldt River basin, Northern Nevada, USGS Bulletin 2210-D, Version 1.0. US Geological Survey, Denver, CO. pubs.usgs.gov/bul/b2210-d/B2210-D-508.pdf

National Research Council (1974) Rehabilitation potential of western coal lands: a report to the energy policy project of the Ford foundation. Ballinger, Cambridge

National Research Council (1979) Surface mining of non-coal minerals: a study of mineral mining from the perspective of the surface mining control and reclamation act of 1977: a report. National Academy of Sciences, Washington, DC

National Research Council (2002) Coal waste impoundments-risks, responses, and alternatives. National Academy Press, Washington, DC

Nelson JD, Abt SR, Volpe RL, van Zyl D, Hinkle HE, Staub WP (1986) Methodologies for evaluating long-term stabilization designs of uranium mill tailing impoundments, NUREG/CR-4620. US Nuclear Regulatory Commission, Washington, DC

Northwest Mining Association (1990) Environmental compliance. Solutions that work, Northwest Mining Association and

Colorado Mining Association seminar, Denver, May 21–23, 1990. Northwest Mining Association, Spokane, WA

OECD (1978) In: Proceedings of the seminar on management, stabilisation and environmental impact of uranium mill tailings, organized by the OECD Nuclear Energy Agency, Albuquerque, United States. Organization for Economic Co-Operation and Development, Paris

OTA (1992) Managing industrial solid wastes from manufacturing, mining, oil and gas production, and utility coal combustion, Background paper OTA-BP-0-82. Congress of the United States, Office of technology assessment, US Government Printing Office, Washington, DC. www.fas.org/ota/reports/9225.pdf

Oweis IS, Khera RP (1998) Geotechnology of waste management, 2nd edn. PWS, Boston

Paşamehmetoğlu AG, Özgenoğlu A (eds) (1998) Environmental issues and management of waste in energy and mineral production. In: Proceedings, 5th international conference, Ankara, Turkey, 18–20 May, 1998. AA Balkema, Rotterdam

PEDCo (1984) Evaluation of management practices for mine solid waste storage, disposal, and treatment. American Mining Congress, Washington, DC

Peterson SD (1990) RCRA's solid waste regulation and its impact on resource recovery in the minerals industry, Mineral issues, An analytical series. US Department of the Interior, Bureau of Mines, Washington, DC

Pusch R (1994) Waste disposal in rock, developments in Geotechnical Engineering, 76. Elsevier Science, Amsterdam

Puura E, Marmo L, D'Alessandro M (eds) (2003) Workshop on mine and quarry waste-the burden from the past. Official Publications of the European Communities, Luxembourg

Ripley EA, Redmann RE, Crowder AA, Ariano TC, Corrigan CA, Farmer RJ, Jackson LM (1996) Environmental effects of mining. Lucie Press, Delray Beach

Ripley EA, Redmann RE, Maxwell J (1982) Environmental impact of mining in Canada. Centre for Resource Studies, Queen's University, Kingston

Ritcey GM (1989) Tailings management: problems and solutions in the mining industry. Elsevier, Amsterdam

Rockwell RB, McDougal RR, Gent CA (2005) Remote sensing for environmental site screening and watershed evaluation in Utah mine lands-East Tintic mountains, Oquirrh mountains, and Tushar mountains, Scientific investigations report 2004-5421. US Geological Survey, Reston, VA. pubs.usgs.gov/sir/2004/5241/pdf/SIR04-5421_508.pdf

Salomons W, Förstner U (eds) (1988) Chemistry and biology of solid waste: dredged material and mine tailings. Springer, Berlin

Salomons W, Förstner U (eds) (1988) Environmental management of solid waste: dredged material and mine tailings. Springer, Berlin

Sarsby RW, Felton AJ (2007) Geotechnical and environmental aspects of waste disposal sites. In: Proceedings of green4 international symposium on geotechnics related to the environment, Wolverhampton, UK, 28 June – July 1, 2004. Taylor & Francis, London

Sarsby R (2000) Environmental geotechnics. Thomas Telford, London

Sengupta M (1993) Environmental impacts of mining: monitoring, restoration, and control. Lewis, Boca Raton

Serne RJ, Peterson SR, Gee GW (1983) Laboratory measurements of contaminant attenuation of uranium mill tailings leachates by sediments and clay liners, technical report NUREG/CR-3124, PNL-4605. Prepared by Pacific Northwest Lab, Richland, WA for US Nuclear Regulatory Commission, Washington, DC

Sherwood DR, Serne RJ (1983) Tailings treatment techniques for uranium mill waste: a review of existing information, technical report NUREG/CR-2938; PNL-4453. Prepared by Pacific Northwest Lab, Richland, WA for US Nuclear Regulatory Commission, Washington, DC

Singhal RK, Mehrotra AK (eds) (2000) Environmental issues and management of waste in energy and mineral production. In: Proceedings 6th international conference, Calgary, AB, 30th May–2nd June, 2000. AA Balkema, Rotterdam

Soderberg RL, Busch RA (1977) Design guide for metal and nonmetal tailings disposal, Information Circular 8755. US Department of the Interior, Bureau of Mines, Washington, DC

Stanley G, Gallagher V, Mhairtin FN, Brogan J, Lally P, Doyle E, Farrell L (2009) Historic mine sites-inventory and risk classification, vol I. A joint study carried out by The Environmental Protection Agency and The Geological Survey of Ireland. www.dcenr.gov.ie/.../New+Collaborative+Study+by+the+EPA+and+GSI.htm

Symonds Group (2001) A Study on the costs of improving the management of mining waste, report to DG environment by Symonds group, in association with COWI, East Grinstead, Great Britain. http://ec.europa.eu/environment/waste/minng/legis.htm

Symposium on management of uranium mill tailings, low-level waste, and hazardous waste (1984) Management of uranium mill tailings, low-level waste and hazardous waste. In: Proceedings of the sixth symposium, February 1–3. Colorado State University, Fort Collins, CO

Symposium on Management of Uranium Mill Tailings, Low-Level Waste, and Hazardous Waste (1985) Management of uranium mill tailings, low-level waste and hazardous waste. In: Proceedings of the seventh symposium, February 1–3. Colorado State University, Fort Collins, CO

Symposium on uranium tailings management (1978) Uranium mill tailings management. In: Proceedings of a symposium, November 20, 21, 1978, organized by the Geotechnical Engineering Program, Civil Engineering Department, Colorado State University, Fort Collins, CO

Szymanski MB (1999) Evaluation of safety of tailings dams. BiTech, Vancouver

International Conference on Tailings & Mine Waste (2004) Proceedings of the eleventh tailings and mine waste conference, 10–13 October 2004, Vail, CO. AA Balkema, Leiden

Technical Committee on Research and Reclamation (2003) symposium Proceedings 1977–2002, CD-ROM, Digital Archive. Richmond, BC

Thorne PD (1992) Information for consideration in reviewing groundwater protection plans for uranium mill tailings sites. US Nuclear Regulatory Commission, Washington, DC

Trivedy RK, Sinha MP (1990) Impact of mining on environment. Ashish Publishing House, New Delhi

Twardowska I, Allen HE, Kettrup AA, Lacy WJ (2004) Solid waste: assessment, monitoring and remediation, vol 4, Waste Management Series. Elsevier, Amsterdam

United Nations Environment Programme (2001) Environmental aspects of phosphate and potash mining. International Fertilizer Industry Association, Paris

UNEP (2001) APELL for Mining, UNEP DTIE. Division of Technology, Industry and Economics, Paris. www.uneptie.org/pc/apell/publications

COLD US (1994) Tailings dam incidents. US Committee on Large Dams, Denver

US Congress Office of Technology Assessment (1992) Managing industrial solid wastes from manufacturing, mining, oil and gas production, and utility coal combustion-background paper, OTA-BP-O-82. US Government Printing Office, Washington, DC

US EPA (1980) Procedures manual for ground water monitoring at solid waste disposal facilities, SW-611, Office of Water & Waste Management. US Environmental Protection Agency, Washington, DC. nepis.epa.gov/.../2000QR04.TXT?...EPA...pubnumber%5E%22530SW611%22

US EPA (1985) Report to congress: wastes from the extraction and beneficiation of metallic ores, phosphate rock, asbestos, overburden from uranium mining, and oil shale, EPA/530-SW-85-033. US Environmental Protection Agency, Office of Solid Waste and Emergency Response, Washington, DC

US EPA (1994) US Environmental Protection Agency Mine Waste Policy Dialogue committee meeting summaries and supporting materials, EPA 530-R-94-043. US Environmental Protection Agency, Office of Solid Waste, Special Waste Branch, Washington, DC

US EPA (1994) Design and evaluation of tailings dams, technical report EPA530-R-94-038. US Environmental Protection Agency, Office of Solid Waste, Special Waste Branch, Washington, DC

US EPA (1994) Extraction and beneficiation of ores and minerals, vol 1: Lead-zinc, technical resource document EPA 530-R-94-011. US Environmental Protection Agency, Office of Solid Waste, Special Waste Branch, Washington, DC

US EPA (1994) Extraction and beneficiation of ores and minerals, vol 2: Gold, technical resource document EPA 530-R-94-013. US Environmental Protection Agency, Office of Solid Waste, Special Waste Branch, Washington, DC

US EPA (1994) Extraction and beneficiation of ores and minerals, vol 3: Iron, technical resource document EPA 530-R-94-030. US Environmental Protection Agency, Office of Solid Waste, Special Waste Branch, Washington, DC

US EPA (1994) Extraction and beneficiation of ores and minerals, vol 4: Copper, technical resource document EPA 530-R-94-031. US Environmental Protection Agency, Office of Solid Waste, Special Waste Branch, Washington, DC

US EPA (1994) Extraction and beneficiation of ores and minerals, vol 5: Uranium, technical resource document EPA 530-R-94-032. US Environmental Protection Agency, Office of Solid Waste, Special Waste Branch, Washington, DC

US EPA (1994) Extraction and beneficiation of ores and minerals, vol 6: Gold placers, technical resource document EPA 530-R-94-035. US Environmental Protection Agency, Office of Solid Waste, Special Waste Branch, Washington, DC

US EPA (1994) Extraction and beneficiation of ores and minerals, vol 7: Phosphate and molybdenum, technical resource document EPA 530-R-94-034. US Environmental Protection Agency, Office of Solid Waste, Special Waste Branch, Washington, DC

US EPA (1994) Innovative methods of managing environmental releases at mine sites, OSW doc 530-R-94-012. US Environmental Protection Agency, Office of Solid Waste, Special Wastes Branch, Washington, DC

US EPA (1994) Treatment of cyanide heap leaches and tailings, technical report EPA 530-R-94-037. US Environmental Protection Agency, Office of Solid Waste, Special Waste Branch, Washington, DC

US EPA (1994) Acid mine drainage prediction, technical report EPA 530-R-94-036, US Environmental Protection Agency, Office of Solid Waste, Special Waste Branch, Washington, DC

US EPA (1995) Human health and environmental damages from mining and mineral processing wastes. US Environmental Protection Agency, Office of Solid Waste, Special Wastes Branch, Washington, DC. http://www.epa.gov/waste/nonhaz/industrial/special/mining/minedock/damage/index.htm. Accessed February 27, 2010

US EPA (2008) Mineral processing waste. US Environmental Protection Agency, Office of Solid Waste, Special Wastes Branch, Washington, DC. http://www.epa.gov/osw/nonhaz/industrial/special/mineral/index.htm. Accessed February 27, 2010

US EPA (2009) Mining waste. US Environmental Protection Agency, Office of Solid Waste, Special Wastes Branch, Washington, DC. http://www.epa.gov/osw/nonhaz/industrial/special/mining/index.htm. Accessed February 27, 2010

US MSHA (2003) Internal review of MSHA's actions at the Big branch refuse impoundment, Martin County Coal Corporation, Inez, Martin County, Kentucky. US Department of Labor, Mine Safety and Health Administration, Arlington, VA. http://www.msha.gov/MEDIA/PRESS/2003/Report20030113.pdf

US NRC (1980) Final generic environmental impact statement on uranium milling, project M-25. Office of Nuclear Material Safety and Safeguards, US Nuclear Regulatory Commission, Washington, DC

van Zyl D (1993) Mine waste disposal. In: Daniel DE (ed) Geotechnical practice for waste disposal, Chapter 12. Chapman & Hall, London, pp 269–286

van Zyl D, Koval M, Li TM (1992) Risk assessment/management issues in the environmental planning of mines. Society for Mining, Metallurgy, and Exploration, Inc, Littleton van Zyl DJA, Vick SG (eds) (1988) Hydraulic fill structures: a specialty conference sponsored by the geotechnical engineering division of the American Society of Civil Engineers (ASCE), co-sponsored by the Society of Mining Engineers of AIME, Colorado State University, Fort Collins, CO, August 15–18, 1988. American Society of Civil Engineers (ASCE), New York

Vick SG (1983) Planning, design, and analysis of tailings dams. Wiley, New York

Walters WH, Skagss RL, Foley MG, Beedlow PA (1986) Erosion protection of uranium tailings impoundments, technical report NUREG/CR-4480, PNL-5724. Prepared by Pacific Northwest Lab, Richmond, WA for the US Nuclear Regulatory Commission, Washington, DC

Waters JA, Brink D (eds) (1987) Mining and industrial waste management. In: Proceedings of the international conference, Johannesburg. South African Institution of Civil Engineers, Johannesburg

Whitby-Costescu L, Shillabeer J, Coates DF (1977) Pit slope manual, Chapter 10-Environmental planning; CANMET (Canadian Centre for Mineral and Energy Technology, formerly Mines Branch, Energy Mines and Resources, Canada, CANMET report 77-2. Ottawa

Williams RE (1975) Waste production and disposal in mining, milling, and metallurgical industries. Miller Freeman, San Francisco

Wilson D (ed) (1981) Design and construction of tailings dams: Proceedings of a seminar, November 6–7, 1980, Colorado School of Mines Press, Golden, CO

Yegulalp TM, Kim K (eds) (1990) Environmental issues and waste management in energy and minerals production. In: Proceedings of the first international conference, Battelle Press, Columbus-Richland, Secaucus, NJ, August, 27–29

Modern Nuclear Fuel Cycles

JAMES S. TULENKO
Laboratory for Development of Advanced Nuclear Fuels and Materials, University of Florida, Gainesville, FL, USA

Article Outline

Glossary

Uranium Occurs in most rocks in concentrations of 2 (sedimentary rocks) to 4 (granite) ppm.

Thorium More readily available nuclear fuel than uranium, being four times more abundant than uranium in the earth's crust.

Uranium dioxide (UO_2) An insoluble oxide of uranium which is the form commonly used in commercial nuclear fuel.

Pitchblende An ore with a very high UO_2 content of up to 70%. Pitchblende also contains radium, thorium, cerium, and lead.

The United States Nuclear Regulatory Commission Regulatory body for radioactive materials and nuclear power plants.

The department of energy Required by law to be responsible for the spent fuel and collects a fee of 1 mill/kWh of nuclear electricity for disposal.

MWD/MTU (mega watt days of energy produced per metric ton of uranium contained) Energy produced per metric ton of uranium (fuel) contained. Current Nuclear Regulatory Commission limits for nuclear power plant fuel is 60,000 MWD/MTU. Normal lifetime of a fuel assembly is 55,000 MWD/MTU.

Definition of the Subject

The Nuclear Fuel Cycle describes the entire process followed to convert uranium or thorium ore to its useful state in nuclear power reactors, and its ultimate and current disposal. The cycle has been followed since the 1960s to produce electrical power safely, and without emissions of environmentally endangering carbon gases.

Introduction

Madame Marie Curie, a pioneer in the field of radioactivity and the first person honored with two Nobel Prizes, discovered radioactivity in 1898, and since then many radioactive elements have seen various uses. One in particular, uranium, has been used for its fissionable properties. When a uranium (or plutonium) atom fissions, it releases approximately 200 MeV of energy. The burning of a carbon (coal) atom releases merely 4 eV. The difference between the two – a 50 million-times advantage in nuclear energy release – shows the tremendous advantage in magnitude between chemical and nuclear energy. This advantage is used for common good in nuclear power reactors around the world. Currently, the United States lags France in the use of nuclear power with the United States obtaining ~20% of its electrical energy from nuclear power plants, while France obtains ~80% of its electrical energy from nuclear power plants. Worldwide, approximately 18% of all electrical energy is produced by nuclear power plants.

The Nuclear Fuel Cycle

The nuclear fuel cycle uses two naturally occurring elements, uranium and thorium, which are both relatively common metals. Both materials are obtained by mining the earth. Uranium occurs in most rocks in concentrations of 2 (sedimentary rocks) to 4 (Granite) ppm. Uranium also occurs in seawater in a concentration of 0.003 ppm, which corresponds to approximately 4 billion tons of uranium in the oceans. Uranium (1.8 g/t) is more abundant than common materials such as silver (0.07 g/t), tungsten (1.5 g/t) and Molybdenum (1.5 g/t) [1]. It is 800 times more abundant than gold. Natural (as mined) uranium contains in atomic abundance 99.2175% Uranium-238 (U-238); 0.72% Uranium-235 (U-235); and 0.0055% Uranium-234 (U-234). Uranium has atomic number 92, meaning all uranium atoms contain 92 protons, with the rest of the mass number being composed of neutrons. All uranium isotopes are radioactive. This radioactive property makes the detection of uranium deposits relatively easy, even allowing for prospecting by air. Uranium-238 has a half-life of 4.5×10^9 years (4.5 billion years), U-235 has a half-life of 7.1×10^8 years (710 million years), and U-234 has a half-life of 2.5×10^5 years (250,000 years). All the U-234 currently present comes from the decay chain of U-238. Uranium-235 is the only fissile isotope available in nature. Uranium can be a fissionable fuel as mined in pressurized heavy water reactors designed by the Atomic Energy of Canada Limited (AECL). These reactors are termed CANDU for CANada Deuterium Uranium. As a by-product of the operation of a nuclear reactor, uranium-238 absorbs a neutron to form, through radioactve decay, the fissile fuel, plutonium-239. Another fissile isotope, uranium-233 comes from the naturally occurring thorium 232 when it captures a neutron. Finally, as part of the plutonium chain, plutonium-241 is also produced. It is important to note that it is the four odd number isotopes: 233, 235, 239, and 241, which are fissionable.

Thorium is an even more readily available nuclear fuel than uranium, being four times more abundant than uranium in the earth's crust. Thorium is the 39th most common element in the earth's crust and is about as common as lead. Thorium is present in the earth's crust with an average concentration of about 9.6 ppm. Thorium must be converted to a fissile fuel in a nuclear reactor by absorption of a neutron, forming through radioactive decay, the fissionable fuel, uranium-233. Thorium has only one naturally occurring isotope, thorium-232, which is radioactive with a half-life of 1.3×10^{10} years. India, which has large thorium deposits, has been a leader in utilizing thorium to breed uranium-233 to serve as a nuclear fuel. Other countries having major deposits of thorium are Australia, Norway, and the United States.

Uranium History

Uranium was discovered in 1789 by Martin Heinrich Klaproth, a German chemist, in the mineral pitchblende, which is primarily a mix of uranium oxides. No one could identify this new material he isolated, so in honor of the planet Uranus that had just been discovered, he called his new material Uranium. Although Klaproth, as well as the rest of the scientific community, believed that the substance he extracted from pitchblende was pure uranium, it was actually uranium dioxide (UO_2). It was not until 1842 that Eugene-Melchoir Peligot, a French chemist, noticed that "pure" uranium reacted oddly with uranium tetrachloride (UCl_4). He then proceeded to isolate pure uranium by heating the uranium dioxide with potassium in a platinum crucible. Radioactivity was first discovered in 1896 when the French scientist Henri Becquerel accidentally placed some uranium salts near some paper-wrapped photographic plates and discovered the natural radioactivity of uranium.

Uranium compounds have long been used for centuries to color glass. Uranium trioxide (UO_3) was used in the manufacture of a distinctive orange Fiestaware dinnerware. In 1938, Otto Hahn (1879–1968), Lise Meitner (1878–1968), and Fritz Strassmann (1902–1980) were the first to recognize that the uranium atom under bombardment by neutrons, actually split, or fissioned.

When a uranium or plutonium atom is fissioned, it releases approximately 200 MeV of energy, while the burning of a carbon (coal) atom releases 4 eV. This difference of 50 million times in energy release shows the tremendous difference in magnitude between chemical and nuclear energy.

Thorium was discovered in 1829 by the Swedish chemist Jons Jacob Berzelius, who named the element after the Thor, the mythical Scandinavian god of war. He also was the first to isolate cerium, selenium, silicon, and zirconium. Thorium and thorium compounds have the properties of having very high melting temperatures. As a result, it was used for high-temperature application such as coatings on tungsten filaments in light bulbs and for high-temperature laboratory equipment. However, its use outside the nuclear fuel cycle has been greatly diminished because of state and federal laws concerning the handling and disposal of radioactive materials. Thorium is found in the minerals monazite and thorianite.

History of Uranium

The earliest recovery of uranium was from pitchblende, an ore with a very high UO_2 content of up to 70%. Pitchblende also contains radium, thorium, cerium, and lead. It is mostly found with deposits that contain phosphates, arsenates, and vanadates. Uranium exists in nature in two valence states, U^{6+} and U^{4+}. These properties are key to the geological distribution of uranium. U^{6+} is soluble in water, but changes to the insoluble U^{4+} in a reducing environment. The occurrence of reducing environments in riverbeds and seas have led to the formation of rich uranium deposits. A rich uranium deposit contains 2% uranium and economic deposits are as low as 0.1%. Once the ore is mined, it is sent to a mill, which is really a chemical plant that extracts the uranium from the ore. The ore arrives via truck and is crushed, leached, and approximately 90–95% of the uranium is recovered through solvent extraction. During the processing a large waste stream called tails is formed, which contains approximately 98–99.9% of the material mined. Because this waste stream or tails contains all the radioactive daughter products of uranium, such as radon and radium, this waste stream must be carefully controlled and stabilized. The tailings pile must have a cover designed to control radiological hazards for a minimum of at least 200 years and designed for 1,000 years, to the greatest extent reasonably achievable. It must also limit radon (^{222}Rn) releases to 20 $pCi/m^2/s$ averaged over the disposal area. The end uranium product of the milling process is U_3O_8, better known as “yellowcake,” because of its color.

Uranium Conversion and Enriching

The U_3O_8 concentrate must be both purified and converted to uranium hexafluoride (UF_6), which is the form required for the enriching process. At the conversion facility, the uranium oxide is combined with anhydrous HF and fluorine gas in a series of chemical reactions to form the chemical compound UF6. The product UF6 is placed into steel cylinders and shipped as a solid to a gaseous diffusion or gaseous centrifuge plant for enrichment. UF6 is a white crystalline solid at room temperature (its triple point is 64°C (147.3°F) and it sublimes at 56.5°C (133.8°F) at atmospheric pressure). The liquid phase only exists under pressures greater than about 1.5 atmospheres and at temperatures above 64°C. At the enrichment plant, the solid uranium hexafluoride (UF_6) from the conversion process is heated in its container until it becomes a gas. The container becomes pressurized as the solid melts UF_6 gas fills the container. The gaseous diffusion process is based on the difference in rates at which the fluorides of U-235 and U-238 diffuse though barriers. The uranium that has penetrated the barrier side is now slightly enriched in U-235 is withdrawn and fed into the next higher enrichment stage, while the slightly depleted material inside the barrier is recycled back into the next lower stage. It takes many hundreds of stages, one after the other, before the UF_6 gas contains enough uranium-235 to be used as an enriched fuel in reactor. Each barrier has millions of holes per square inch with each hole approximately 10^{-7} in. in diameter. This gaseous diffusion enrichment process is very energy intensive, as the gas is compressed and expanded at each stage.

The other commercial enriching process, which uses an order of magnitude less energy, is the gaseous centrifuge process. The gas centrifuge uranium enrichment process uses a large number of rotating cylinders in series and parallel formations. Centrifuge machines are interconnected to form trains and cascades. In this process, UF_6 gas is placed in a cylinder and rotated at a high speed. This rotation creates a strong centrifugal force so that the heavier gas molecules (containing U-238) move toward the outside of the cylinder and the lighter gas molecules (containing U-235) collect closer to the center. The stream that is slightly enriched in U-235 is withdrawn and fed into the next higher stage, while the slightly depleted stream is recycled back into the next lower stage. At each stage of the gaseous diffusion process the U-235 is enriched by a factor of 1.004, where at each stage of the gaseous centrifuge process the stage enrichment factor is 1.2. For 1 kg of uranium enriched to 5% U-235, 9.4 kg of natural uranium feed are required and 8.4 kg of depleted uranium (tails) with a U-235 isotope content of approximately 0.2% are produced as a waste stream. The US Nuclear Regulatory Commission has decided that depleted uranium is a low level waste. The Department of Energy has over 560,000 mt stockpile of uranium tails stored as UF_6 in steel cylinders. The tails uranium has minor uses as a shields for radioactive sources, as the penetrator in armor piercing shells, as a yacht hold ballast, and as a weight for the balancing of helicopter rotor tips and passenger aircraft.

Nuclear Fuel Fabrication

The enriched UF_6 is transported to a fuel fabrication plant where the UF_6, in solid form in containers, is again heated to its gaseous form, and the UF_6 gas is chemically processed to form uranium dioxide (UO_2) powder. This powder is then pressed into pellets, sintered into ceramic form, loaded into Zircaloy tubes, pressurized with helium and sealed. The fuel rods are then placed into an array (17 × 17) which is bound together with guide tubes, spacer grids and top and bottom end fittings, all of which forms the nuclear fuel assembly. Depending on the type of light water reactor, a fuel assembly may contain up to 264 fuel rods and have dimensions of 5–6 in. square by about 12 ft long. The fuel is placed into containers and is trucked to the nuclear fuel plants to generate electricity. A single pressurized water fuel assembly contains about 500 kg of enriched uranium and can produces 200,000,000 kWh of electricity. Since the average national electrical yearly use per person is 11,867 kWh, a single nuclear fuel assembly gives 5,562 people their yearly electric needs during its 3 years of operation.

Nuclear Fuel Operation and Disposal

Every 12–24 months, US nuclear power plants are shut down and the oldest fuel assemblies are removed (approximately ⅓–½) and replaced with new fuel assemblies. The power production of a fuel assembly is measured in MWD/MTU or mega watt days of energy produced per metric ton of uranium (fuel) contained. Currently the normal lifetime of a fuel assembly is 55,000 MWD/MTU and the maximum lifetime currently allowed by the Nuclear regulatory commission is 60,000 MWD/MTU. At the end of its useful life, the spent fuel assembly is placed in a cooled borated water storage pond to allow for removal of the radioactive decay heat. After approximately 5 years of wet storage, the decay heat has been sufficiently decreased that the fuel assembly can be removed to dry storage in concrete or steel containers. Since only approximately 5% of the uranium fuel is destroyed, in Europe and Asia the spent fuel is reprocessed and the 95% of uranium remaining is recycled, with the 5% of radioactive waste products sent to waste storage. At the current time, the United States policy was to store the spent fuel in a waste repository being built at Yucca Mountain in Nevada. Most recently this site has become part of a political struggle and the current administration has moved to halt all licensing of the Yucca Mountain site and to convene a high level committee to revisit the question of nuclear waste disposal. The Department of Energy is required by law to be responsible for the spent fuel and collects a fee of 1 mill/kWh of nuclear electricity delivered, which is paid by consumers of nuclear-generated electricity. The one assembly described above would generate approximately $200,000 in the waste fund for its disposal.

There is enough uranium and thorium in the world to produce the required amount of fuel to allow nuclear plants to produce the current rate of electrical energy usage for the next 1,000 years.

Future Directions

(A discussion including potential impacts on the development of certain areas of science.) With the dawn of the environmental awareness and new economies of energy production, the nuclear fuel cycle also is undergoing change. Research efforts are continuing to find new and more efficient ways to use the fissionable atom. Also, currently operating nuclear plants are becoming more efficient and cost beneficial. With no greenhouse gases to speak of, nuclear energy is bound to play a role in the nation's future energy needs. More than 100 nuclear reactors nationwide now provide almost 20% of our energy production. The large, proven-safe nuclear plants produce electricity best when running at full power, 24/7.

Bibliography

1. Cochran RG, Tsoulfanidis N (1999) The nuclear fuel cycle: analysis and management. American Nuclear Society, La Grange Park
2. Wilson PD (1996) The nuclear fuel cycle: from ore to waste. Oxford Science Publications, Oxford

Molecular Breeding Platforms in World Agriculture

Jean-Marcel Ribaut, Xavier Delannay, Graham McLaren, Frederick Okono
Integrated Breeding Platform, Consultative Group on International Agricultural Research, Generation Challenge Programme, c/o CIMMYT, Texcoco, Edo. Mexico

Article Outline

Glossary

Analytical pipeline A sequence of data management and statistical analysis algorithms which can be applied to one or more data sets to produce a result which can be interpreted and applied in decision making.

Capacity building Assistance that is provided to entities, usually institutions in developing countries, which have a need to develop a certain skill or competence, or for general upgrading of capability.

Cyberinfrastructure (CI) Computer-based research environments that support advanced data acquisition, data storage, data management, data integration, data mining, data visualization, and other computing and information processing services over the Internet. In scientific usage, CI is a technological solution to the problem of efficiently connecting data, computers, and people with the goal of enabling derivation of novel scientific theories and knowledge.

Gene Segment of DNA specifying a unit of genetic information; an ordered sequence of nucleotide base pairs that produce a certain product that has a specific function.

Information system (IS) An integrated set of computing components and human activities for collecting, storing, processing, and communicating information.

Integrated breeding platform (IBP) Term to describe a Molecular Breeding Platform (see below) in a broader sense including the availability of tools and services suitable for conventional breeding based on phenotypic selection only.

Molecular breeding (MB) Identification, evaluation, and stacking of useful alleles for agronomic traits of importance using molecular markers (MMs) in breeding programs. MB encompasses several modern breeding strategies, such as marker-assisted selection (MAS), marker-assisted backcrossing (MABC), marker-assisted recurrent selection (MARS), and genome-wide selection (GWS).

Molecular breeding platform (MBP) A term that has come to indicate a virtual platform driven by modern information and communication technologies through which MB programs can access genomic resources, advanced laboratory services, and analytical and data management tools to accelerate variety development using marker technologies.

Plant breeding The science of improving the genetic makeup of plants in order to increase their value. Increased crop yield is the primary aim of most plant breeding programs; benefits of the hybrids and new varieties developed include adaptation to new agricultural areas, greater resistance to disease and insects, greater yield of useful parts, better nutritional content of edible parts, and greater physiological efficiency especially under abiotic stress conditions.

Quantitative trait locus (QTL) A region of the genome that contains genes affecting a quantitative trait. Though not necessarily genes themselves, QTLs are stretches of DNA that are closely linked to the genes that underlie the corresponding trait.

Definition of the Subject

In the last decade, private seed companies have benefitted immensely from molecular breeding (MB) [1]. A private sector-led "gene revolution" has boosted crop adaptation and productivity in developed countries, by applying and combining the latest advances in molecular biology with cutting-edge information and communication technologies combined with accurate plant phenotyping.

MB allows the stacking of favorable alleles, or genomic regions, for target traits in a desired genetic background thanks to the use of polymorphic molecular markers (MMs) that monitor differences in genomic composition among cultivars, or genotypes, at specific genomic regions, or genes, involved in the expression of those target traits. The use of MMs generally increases the genetic gain per crop cycle compared to selection based on plant phenotyping only, and therefore reduces the number of needed selection cycles, hastening the delivery of improved crop varieties to the farmers.

In contrast to the private sector, MB adoption is still limited in the public sector, and is hardly used at all in developing countries. This is the result of several factors, among which are the following: (1) scientists from the academic world are more interested in discovering new genes or QTLs to be published than in applied biology; (2) until recently access to genomic resources was limited in the public sector, especially for less-studied crops; (3) public access to large-scale genotyping facilities was not easily available; and (4) although a broad set of stand-alone tools are available to conduct the multiple types of analyses necessitated by MB, no single analytical pipeline is available today in the public sector allowing integrated analysis in a user-friendly mode.

The situation is even more critical in developing countries as additional limitations include shortage of well-trained personnel, inadequate laboratory and field infrastructure, lack of ISs with applicable and flexible analysis tools, as well as inappropriate funding – simply put, resource-limited breeding programs. As a result, the developing world has yet to benefit from the MB revolution, and most of the countries indeed lack the fundamental prerequisites for a move to informatics powered breeding.

Under those circumstances, developing and deploying a sustainable web-based Molecular Breeding Platform (MBP) as a one-stop shop for information, analytical tools, and related services to help design and conduct marker-assisted breeding experiments in the most efficient way will alleviate many of the bottlenecks mentioned earlier. Such a platform will enable breeding programs in the public and private sectors in developing countries to accelerate variety development using marker technologies for different breeding purposes: major genes or transgene introgression via marker-assisted backcrossing (MABC), gene pyramiding via marker-assisted selection (MAS), marker-assisted recurrent selection (MARS) and, in a not too distant future, genome-wide selection (GWS).

Introduction

Since the dawn of agriculture, mankind has sought to improve crops by selecting individual plants with the most desirable characteristics or traits. Agricultural productivity has been progressively enhanced by constant innovation, including improved crop varieties to increase production in specific environments [2]. The major objective of crop improvement is to identify within heterogeneous materials those individuals for which favorable alleles are present at the highest proportion of loci involved in the expression of key traits [3]. The classical plant breeding method is based on increasing the probability of selecting such individuals from populations generated from sexual matings. Selection has traditionally been carried out at the whole-plant level (i.e., phenotype), which represents the net result of genotype and environment (and their interactions). Phenotypic selection has delivered tremendous genetic gains in most cultivated crop species, but is severely limited when faced with traits that are heavily modulated by the environment [4]. In addition, the nature of some traits can make the phenotypic testing procedure itself complex, unreliable, or expensive (or a combination of these).

The recent remarkable development of molecular genetics and associated technologies represents a quantum leap in our understanding of the underlying genetics of important traits for crop improvement. The ongoing revolutions in molecular biology and information technology offer tremendous and unprecedented opportunities for enhancing the effectiveness and efficiency of MB programs. Indirect selection, based on genetic markers, presents an efficient complementary breeding tool to phenotypic selection. Individual genes or QTLs having an impact upon target traits can be identified and linked with one or more markers, and then the marker loci can be used as a surrogate for the trait, resulting in greatly enhanced breeding efficiency [5–8].

Molecular techniques can have an impact upon every stage of the breeding process from parental selection and cross prediction [9], to introgression of known genes [10] and population enhancement. Selection of beneficial alleles of known genes can be done through marker-assisted selection (MAS) – the selection of specific alleles for traits conditioned by a few loci [10] – or through marker-assisted backcrossing (MABC) – transferring specific alleles of a limited number of loci from one genetic background to another, including transgenes [11, 12]. For marker-assisted population improvement, individuals selected from a segregating population based on their marker genotype are intermated at random to produce the following generation, at which point the same process can be repeated a number of times [13]. A second approach aims at direct recombination between selected individuals as part of a breeding scheme, seeking to generate an ideal genotype or ideotype [14]. The ideotype is predefined on the basis of QTL mapping within the segregating population, combined with the use of multi-trait selection indices that can also consider historical QTL data. This variety development approach is commonly referred to as marker-assisted recurrent selection (MARS) [15–17], or genotype construction. An alternative is to infer a predictive function using all available markers jointly, without significant testing and without identifying a priori a subset of markers associated with the traits of interest. This more recent approach coming from genomic medicine [18, 19], and then applied successfully in animal breeding [20] named genome-wide selection (GWS), also appears to be quite promising in crop improvement [7].

Concomitantly with the evolution of marker technologies becoming increasingly “data rich,” the amount of data produced by plant breeding programs has increased dramatically in recent years. Increasingly, the critical factor determining the rate of progress in plant breeding programs is their capacity to manage large amounts of data efficiently and subsequently maximize the timely extraction of meaningful information from that data for use in selection decisions. If genotyping has become less of an issue, the efficient management of genotyping data in a broad sense, including sequence information, is increasingly becoming a major challenge in modern plant breeding. This was recognized early on in the private sector where the establishment of platforms or pipelines integrating field and laboratory processes with powerful data management systems (DMS) that merged and analyzed the data collected at every step and guided the process of crop improvement toward the release of improved cultivars has been the key to successful adoption of MB.

A few initiatives have taken place in the public sector to establish efficient data management or ISs [21, 22]. One of these has been led by several centers of the Consultative Group on International Agricultural Research (CGIAR) which have worked over the past decade, along with advanced research institutes (ARIs) and national agricultural research systems (NARS) in developing countries, to develop an open-source generic IS, the International Crop Information System (ICIS), to handle pedigree information, genetic resource, and crop improvement information [23]. Based on some elements of ICIS, the CGIAR Generation Challenge Programme (GCP, http://www.generationcp.org) has invested in integrating crop information with genomic and genetic information and in using existing or developing new public decision-support tools to access and analyze information resources in an integrated and user-friendly way [24]. Another initiative has been led by Primary Industries and Fisheries (PI&F) of the Queensland Government Department of Employment, Economic Development and Innovation in Australia, which recognized that effective data management is an essential element in obtaining maximum benefit from their investment in plant breeding. In conjunction with the New South Wales Department of Primary Industries (NSW DPI) and more recently Dart Pty Ltd (http://www.diversityarrays.com/) they are in the process of developing a linked IS for plant breeding (Katmandoo) that includes applications for capturing field data using hand-held computers, barcode-based seed management systems, and databases to store and link field trial data, laboratory data, genealogical data, and marker data [25].

Although an IS involves far more than a database, the development and implementation of a suitable database system alone remains a real challenge because of the fast turnover in technologies, the need to manage and integrate increasingly diverse and complex data types, and the exponential increase in data volume. Previous solutions, such as central databases, journal-based publication, and manually intensive data curation, are now being enhanced with new systems for federated databases, database publication, and more automated management of data flows and quality control. Along with emerging technologies that enhance connectivity and data retrieval, these advances should help create a powerful knowledge environment for genotype–phenotype information [26].

In addition to efficient data management, advances in statistical methodology [27–29], graphical visualization tools, and simulation modeling [9, 30–32] have greatly enhanced these ISs. The availability of molecular data linked to computable pedigrees [33] and phenotypic evaluation now makes genotype–phenotype analysis a practical reality [34].

In order to realize the full potential of marker technologies and bioinformatics in plant breeding, tools for molecular characterization, accurate phenotyping, efficient ISs, and effective data analysis must be integrated with breeding workflows managing pedigree, phenotypic, genotypic, and adaptation data. The goals of this integration of technologies are to (1) create genotype–phenotype trait knowledge for breeding objectives, and (2) use that knowledge in product development and deployment [4].

This entry generally explores the pace of innovation in world agriculture and the rise of MB. It particularly illustrates the accelerating application of information and communication technologies to the information management challenges of MB and, as a result, the emergence of virtual molecular breeding platforms (MBPs) as a vital tool for accelerating genetic gains and rapidly developing more resilient and more productive cultivars.

This entry reviews the rationale for access to MB technology and services and the status of existing public analytical pipelines and ISs for MB, and offers a detailed case study for the CGIAR GCP Integrated Breeding Platform (IBP) – the pioneer public sector MBP specifically targeting developing country breeding programs. It explores the gaps between countries and between crops in the application of informatics-powered MB approaches, and the potential for adopting MBPs to close these gaps; and it reviews institutional, governmental, and public support for these approaches. The entry discusses the challenges and opportunities inherent in MBPs, and the potential economic impact of MB. Finally, the entry explores the future directions and perspectives of MBPs.

Marker Technologies and Service Laboratories

Markers are “characters” whose pattern of inheritance can be followed at the morphological (e.g., flower color), biochemical (e.g., proteins and/or isozymes), or molecular (DNA) levels. They are so called because they can be used to elicit, albeit indirectly, information concerning the inheritance of "real" traits. The major advantages of molecular over other classes of markers are that their number is potentially unlimited, their dispersion across the genome is complete, their expression is unaffected by the environment and their assessment is independent of the stage of plant development [35]. During the past two decades, DNA technology has been exploited to advance the identification, mapping, and isolation of genes in a wide range of crop species. The first generation of DNA markers, restriction fragment length polymorphisms (RFLPs), was used to construct the earliest genome-wide linkage maps [36] and identify the first QTLs [37, 38]. During the 1990s, emphasis switched to assays based on the polymerase chain reaction (PCR), which are much easier to use and potentially automatable [39]. The development of simple sequence repeats (SSRs) [40], amplified fragment length polymorphisms (AFLPs) [41], and single nucleotide polymorphism (SNP) [42] opened the door for large-scale deployment of marker technology in genomics and progeny screening.

SNPs are amenable to very high throughput and a wide range of detection techniques has been developed for them, from singleplex systems to high-density arrays. They can be used in fully integrated robotic systems going from automated DNA extraction to automated scoring in high-throughput detection platforms. The combination of increase in throughput and lowering in costs makes SNPs highly suitable to intensive marker applications in plant breeding such as MARS and the emerging approach of GWS. Based on SNP technology, production of molecular marker (MM) data expanded more than 40-fold between 2000 and 2006 at Monsanto, while cost per data point decreased to one sixth of the original cost [43].

With the transition from SSRs to SNPs and the concomitant large increase in the demand for genotyping as markers get more and more widely used in a broad range of applications from medicine to plant breeding, marker genotyping laboratories have evolved from relatively low-tech operations to highly automated, high-throughput laboratories using an array of sophisticated equipment (pipetting robots, high-density PCR, high-throughput SNP detection machines, high-level informatics). Although large private seed companies have had the need and the resources to put in place large-scale genotyping laboratories for their own uses, smaller programs, especially in the public sector, have typically not had the resources or the justification to establish such large operations to respond to their increasing need for SNP genotyping data. In response to this need, a few private marker service laboratories have sprung up over the past few years, which can provide complete genotyping services for their customers, from DNA extraction to generation of large numbers of SNP or other datapoints. Due to their broad customer base (from medical research laboratories to animal and plant breeding operations, both public and private), these laboratories can have a large volume of datapoint production which may lead to low costs for the customer and high throughput. They are able to invest in the most advanced equipment to keep up with the constant evolution of genotyping technologies and are able to pass on the resulting benefits to their customers. Processes have now been put in place for rapid shipment of leaf samples from any location (field or laboratory) around the world without any restrictions. Examples of such companies that can service breeding programs from around the world are DNA LandMarks, Inc. of Saint-Jean-sur-Richelieu, Quebec, Canada (http://www.dnalandmarks.ca/english/) and KBioscience Ltd. of Hoddesdon Herts, UK (http://www.kbioscience.co.uk/). For many public breeding programs and small companies, especially in developing countries, it is now more efficient to use those types of contract genotyping services than to try to support their growing MB needs through the establishment of an in-house laboratory. Functional and reliable SNP laboratories are especially difficult to establish in many developing countries due to the unreliability of the power supply, difficulties in shipping and storing and a low level of resources for the purchase and maintenance of sophisticated equipment. The GCP is facilitating the linkage between users and service laboratories through its marker services, a component of the breeding services offered through the GCP's IBP.

Analytical Tools, Software, and Pipelines

One of the achievements of the plant biotechnology revolution of the last two decades has been the development of molecular genetics and associated technologies, which have led to the development of an improved understanding of the basis of inheritance of agronomic traits. The genomic segments or QTLs involved in the determination of phenotype can be identified from the analysis of phenotypic data in conjunction with allelic segregation at loci distributed throughout the genome. Because of this, the mode of inheritance, as well as the gene action underlying the QTL, can be deduced [44]. As with the improvement in marker technologies, the statistical tools needed for QTL mapping have evolved from a rudimentary to a very sophisticated level [45]. Previous approaches based on multiple regression methods, using least squares or generalized least squares estimation methods [46, 47], have evolved to composite interval mapping [9], mixed model approaches using maximum likelihood or restricted maximum likelihood (REML) [48], and Markov Chain Monte Carlo (MCMC) algorithms [49, 50], which use Bayesian statistics to estimate posterior probabilities by sampling from the data. In parallel, with progress in the characterization of genetic effects at QTLs and refinement of QTL peak position through meta-analysis [51], advances have also been made in understanding the impact of the environment on plant phenotype. The mapping of QTLs for multiple traits has allowed the quantification of QTL by environment interaction (QEI) [52] and, more recently, approaches using factorial regression mixed models have been applied to model both genotype by environment interaction [53] and QEI [48, 54, 55]. Recent approaches are now implemented to evaluate gene networking [56] and epistasis, based on Bayesian approaches [57, 58] or through stepwise regression by considering all marker information simultaneously [59, 60]. Epistasis and balanced polymorphism influence complex trait variation [61, 62], and classical generation means analyses, estimates of variance components, and QTL mapping indicated an important role of digenic and/or higher-order epistatic effects for all biomass-related traits in model plants [63] and in crops [64–66]. It will be critical to implement the most efficient MB strategies in order to evaluate and include these genetic effects in breeding schemes [60].

All tools necessary to run MB projects, from the simplest to the most complicated approaches, are available today in the public domain. They are based on different algorithms and statistical approaches, from the very simple to the more complex. One challenge is the diversity of tools available for a given analytical function or along the different steps of an analytical pathway, making the choice of the "right" tool difficult and the move from one analytical step to the next very tedious due to the complete lack of common standards and formatting across tools. The number of applications available for QTL analysis illustrates well the multiplicity and diversity of tools that are available for a given analysis. The following software packages have been developed over the past 20 years:

- Mapmaker/QTL [67]
- MapQTL [68, 69]
- QTL Cartographer [9, 70]
- PLABQTL [71]
- QGene [72, 73]
- Map Manager QT [74]
- iCIM [59, 60]

For most of these applications, the first versions were already available 15 years ago and the multiplicity and possible duplication generated by the independent development of these tools were already identified at the Gordon Research Conference on Quantitative Genetics and Biotechnology held in February 1997 in Ventura, California. A main objective of that workshop was to survey participants on the attributes of several software packages for QTL mapping and to define their analytical needs which were not presently met by the existing software packages. The workshop covered software for QTL mapping in inbred and outcrossed populations and the conclusions are available at: http://www.stat.wisc.edu/~yandell/statgen/software/biosci/qtl.html. In those conclusions one can read that "[a] consensus was reached that there is considerable overlap in the kinds of matings handled and statistics produced by the various QTL mapping software packages," clearly identifying the need for better coordinated efforts. Such coordination never took place, as is often the case in public research. As a result, most of those QTL packages are still available today, although in more sophisticated versions. They are all suitable for QTL mapping but use different statistical algorithms, present a different user interface, and necessitate different input and output file formats.

Some specialists in the field realized that the public software packages are usually too specialized and too technical in statistics to permit a thorough understanding by the many experimental geneticists and molecular biologists who would want to use them. In addition, the fast methodological advances, coupled with a range of stand-alone software, make it difficult for expert as well as non-expert users to decide on the best tools when designing and analyzing their genetic studies. Based on this rationale, a few commercial analytical pipelines emerged about a decade ago that include some of the QTL packages mentioned above. Two of them are Kyazma and GenStat®. These applications assist plant scientists by providing easy access to statistical packages for phenotypic and genotypic data. Kyazma was founded in the spring of 2003 (http://www.kyazma.nl/), and offers powerful methods for genetic linkage mapping and QTL analysis. Since 2003 Kyazma has taken over the development of the software packages JoinMap® and MapQTL® from Biometris of Plant Research International. Kyazma handles the distribution and support of JoinMap and MapQTL and, in collaboration with the statistical geneticists of Biometris, Kyazma provides introductory courses on genetic linkage mapping and QTL analysis in order to make the use of the software even more accessible. GenStat encompasses statistical data analysis software for biological and life science markets worldwide. GenStat includes the ASReml algorithm (average information algorithm for REML) to undertake very efficient meta-analyses of data with linear mixed models. The development of GenStat at Rothamsted began in 1968, when John Nelder took over from Frank Yates as Head of Statistics. Roger Payne took over leadership of the GenStat activity when John Nelder retired in 1985 (http://www.vsni.co.uk/). An important feature of GenStat is that it has been developed in (and now in collaboration with) a Statistics Department whose members have been responsible for many of the most widely used methods in applied statistics. Examples include analysis of variance, design of experiments, maximum likelihood, generalized linear models, canonical variates analysis, and recent developments in the analysis of mixed models by REML.

These commercial analytical pipelines offer a set of quality tools to researchers in plant science. However, they cover only a part of the configurable workflow system that is required for integrated breeding activities. In addition, there is a need to have tools and analytical pipelines that are freely available and, if possible, based on open source code to avoid dependence on private companies that might discontinue support and ensure access to the tools even with limited financial resources, which is a critical constraint in the arena of research for development, of which breeding programs of developing countries are key partners. It is important to underline that a version of GenStat that does not include the most advanced version of the different tools but allows users to run most basic analyses is available for breeding programs in developing countries. The web site for the GenStat Discovery Edition is http://www.vsni.co.uk/software/genstat-discovery/, but this version of the pipeline does not include QTL selection based on the mixed model approach, which is available in the commercial version.

The issue of open source code is an important one as, even for freely-available tools, the lack of availability of the source code limits the further expansion and customization of the tools. It also reduces the opportunity of researchers in developing countries to participate in methodology development. Over the last decade, a programming language and software environment for statistical computing and graphics, R, is becoming the reference in open source code for a broad range of biological applications, including genetic analysis (http://www.r-project.org/). Its source code is freely available under the *GNU General Public License* (http://en.wikipedia.org/wiki/GNU_General_Public_License). The R language has become a de facto standard among statisticians for the development of statistical software. It compiles and runs on a wide variety of UNIX, Windows, and MacOS platforms. R is similar to other programming languages, such as C, Java, and Perl, in that it helps people perform a wide variety of computing tasks by giving them access to various commands. For statisticians, however, R is particularly useful because it contains a number of built-in modules for organizing data, running calculations on the information, and creating graphical representations of the data sets. R provides a wide variety of statistical (linear and nonlinear modeling, classical statistical tests, time-series analysis, classification, clustering, etc.) [29] and graphical techniques, and is highly extensible. Close to 1,600 different packages reside on just one of the many web sites devoted to R, and the number of packages has grown exponentially. However, R is difficult to use directly and procedures based on R must be wrapped in user-friendly menu systems if field biologists are to use them.

Information Systems

A functional IS involves far more than an analytical pipeline; it is a complete system that should include:

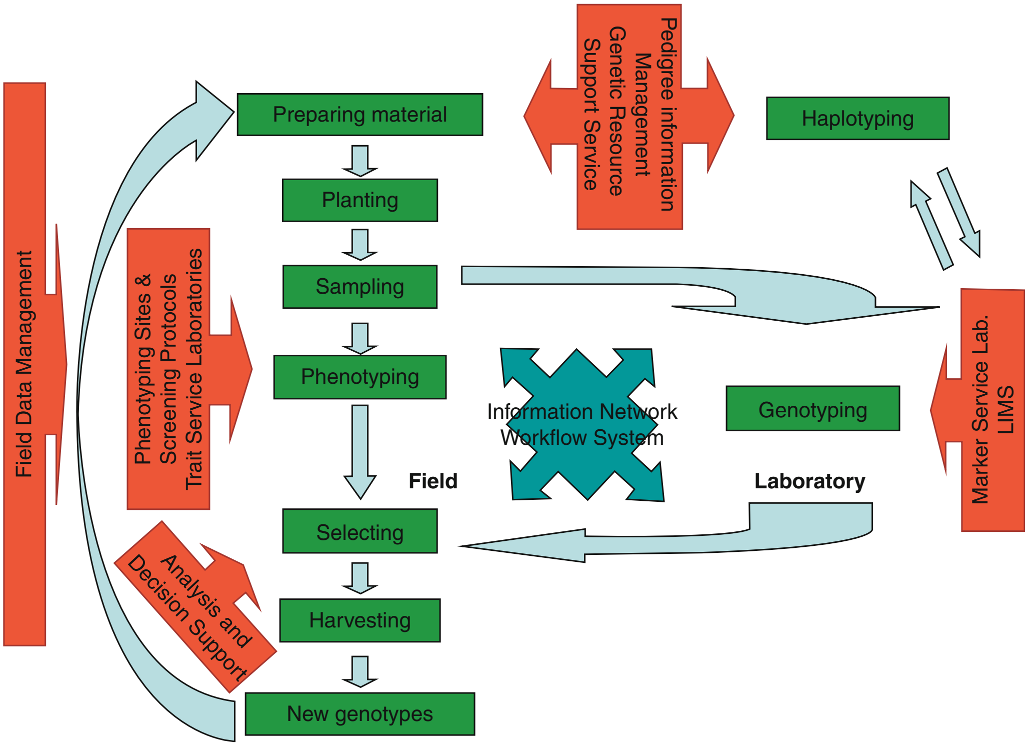
- A project planning module
- A germplasm management module
- A robust relational database
- Analytical standards
- Data collection and cleaning tools
- Analytical and decision support tools
- Query tools
- A cyber infrastructure (CI) that links the different tools in a cohesive and user-friendly way

Key elements of an IS are obviously the CI and the DMS as described in the following section. The value of an IS does not only reside in the quality of the individual tools or modules that are part of it, but rather in the CI or middleware that ensures cohesion across tools and efficient communication with databases.

There are not many examples of breeding ISs in the public domain. One example is the ICIS (http://www.icis.cgiar.org, [23]). ICIS is an open source IS for managing genetic resource and breeding information for any crop species. It has been developed over the last 10 years through collaboration between centers of the CGIAR, some NARS, and private companies. The ICIS system is Windows-based, and distributable on CD-ROM or via the Internet. It contains a genealogy management system (GMS, [33]) to capture and process historical genealogies as well as to maintain evolving pedigrees and to provide the basis for unique identification using internationally accepted nomenclature conventions for each crop; a seed inventory management system (IMS); a DMS [75] for genetic, phenotypic, and environmental data generated through evaluation and testing, as well as for providing links to genomic maps; links to geographic ISs that can manipulate all data associated with latitude and longitude (e.g., international, regional, and national testing programs); applications for maintaining, updating, and correcting genealogy records and tracking changes and updates; applications for producing field books and managing sets of breeding material, and for diagnostics such as coefficients of parentage and genetic profiles for planning crosses; tools to add new breeding methods, new data fields, and new traits; and tools for submitting data to crop curators and for distributing data updates via CD-ROM and electronic networks. The community of ICIS collaborators communicates via the ICIS Wiki (http://www.icis.cgiar.org), where all design and development decisions are documented. Feature requests and bug reports are made through the ICIS Communications project and the source code is published through various other ICIS projects on CropForge (http://cropforge.org). A commercial company, Phenome-Networks, has implemented a Web-based IS based on ICIS (http://phnserver.phenome-networks.com/).

Another system available is the Katmandoo Biosciences Data Management System (http://www.katmandoo.org/, [25]), which is a freely available, open source DMS for plant breeders developed by PI&F, NSW DPI, and DArT Pty. Ltd. It comprises linked ISs for plant breeding including applications for capturing field data using hand-held computers, barcode-based seed management systems, and databases to store and link field trial data, laboratory data, genealogical data, and marker data. A particular focus is on the use of whole-genome MM information to create graphical genotypes, track the ancestral origin of chromosomal regions, validate pedigrees, and infer missing data. It includes the applications of the Pedigree-Based Marker-Assisted Selection System (PBMASS) developed by PI&F as well as a seed management system, a digital field book for hand-held computers, and a system for directly recording weights of barcoded samples.

Both ISs struggle with the problem of integrating the different components into a single configurable system which matches the workflows of different breeding projects. Such a workflow should provide the user all tools and analytical means required to run a crop cycle: from germplasm preparation and planting, through the collection of phenotypic and the production of the genotypic data and their analysis, to the identification of genotypes to be



Molecular Breeding Platforms in World Agriculture. Figure 1
Different activities conducted during the crop cycle of an MB experiment presented in a generic way

crossed or the selection of suitable genotypes to be planted in the next cycle (Fig. 1).

In order to do this effectively, a CI is required which allows syntactic linkage between different data resources and applications.

Cyberinfrastructure and Data Management

We have referred to the revolution in Information and Communication Technology and the opportunities it presents for improving the efficiency of plant breeding. However, plant breeding is not the only area of biology being affected by this revolution and, in fact, the successful deployment of MB depends on other fields of information-intensive biology delivering knowledge (markers and methodology) to plant breeding. Even more is expected of the information and communications technology (ICT) revolution in the developing world, as it offers an opportunity for scientists there to overcome some of the constraints of isolation, the "brain drain," and the lack of infrastructure which have prevented them from fully participating in science for development in the past [76].

It is generally recognized that upstream biology is increasingly reliant on networks of integrated information and on applications for analyzing and visualizing that information. Discipline-specific (sequence and protein databases) and model organism ISs such as Graingenes (http://wheat.pw.usda.gov/GG2/index.shtml), Gramene (http://www.gramene.org/), MaizeGDB (http://www.maizegdb.org/), and Soybase (http://www.soybase.org/) have been developed to facilitate exchanges in molecular biology and functional genomics. As noted above, plant breeding depends on these upstream sciences of molecular biology, functional genomics, and comparative biology to deliver the knowledge needed to deploy MB. The bottleneck in the overall network has been the technology needed to integrate diverse and distributed information resources, and many information scientists have been working on this problem [24, 26, 77].

One constraint to integration of scientific information is the necessity to have a standard terminology for biological concepts across species and disciplines. A successful example of such standardization is the Gene Ontology (GO) initiative (http://www.geneontology.org, [78]). Another more specialized ontology initiative, especially pertinent to agriculture, is the Plant Ontology Consortium (POC: http://www.plantontology.org, [79–81]). However, these formal descriptions remain somewhat limited to biology of model plants and controlled environments. A key challenge will be to extend such standards to describe characteristics of plants growing in the unique, stress-prone environments found within the developing world to ensure a wider impact of such standards on international agriculture. The GCP has been working with POC to expand these ontologies to economic traits and farming environments so that they can be used in the field of plant breeding [82].

Another constraint to the efficient utilization of genomic information is the sheer volume of sequence data that can now be generated very cheaply across numerous genotypes. ISs to handle this volume of information are struggling to keep up. In plant biology, some examples of systems aiming to handle these torrents of data are the Germinate database ([83], http://bioinf.scri.ac.uk/public/?page_id=159) and the Genomic Diversity and Phenotype Connection (GDPC, http://www.maizegenetics.net/gdpc/). The primary goal of Germinate is to develop a robust database which may be used for the storage and retrieval of a wide variety of data types for a broad range of plant species. Germinate focuses on genotypic, phenotypic, and passport data, but has been designed to potentially handle a much wider range of data including, but not limited to, ecogeographic, genetic diversity, pedigree, and trait data, and will permit users to query across these different types of data. The developers have aimed to provide a versatile database structure, which can be simple, requires little maintenance, may be run on a desktop computer, and yet has the potential to be scaled to a large, well-curated database running on a server. The design of Germinate provides a generic database framework from which interfaces ranging from simple to complex may be used as a gateway to the data. The data tables are structured in a way that they are able to hold information ranging from simple data associated with a single accession or plant, to complex data sets, images, and detailed text information. Features of the Germinate database structure include its ability to access any information associated with a group of accessions and to relate different types of information through their association with an accession. The GDPC database was designed as a research database to support association genetics applications such as Tassel (http://www.maizegenetics.net/index.php?option=com_content&task=view&id=89&Itemid=119) and is being extended to handle higher and higher densities of genotyping and sequence data. The second version of Germinate seems quite similar to GDPC and if new databases are developed to handle the large data files to be generated soon through high-throughput sequencing, some conversion tools should be easily developed to migrate data from one system to another.

Finally, the problem of integrating all these diverse and widely-distributed information resources is a major informatics challenge, which is being tackled on several fronts at several levels of complexity. The BioMOBY project ([84], http://www.biomoby.org, [85]) and the Semantic Web seek to define standards that will allow computer programs to interpret requests for information or services, find informatics resources capable of fulfilling those requests, and return the results without the authors of the interacting software having specifically collaborated. In the private sector, solutions have been more pragmatic and Enterprise Software solutions have been developed to link data resources and applications with specific services. The iPlant Collaborative (http://www.iplantcollaborative.org/) is a National Science Foundation (NSF)-funded initiative designed to bring these Enterprise Software solutions to the biological sciences in the form of CI which can support any biological data resource and analytical application. iPlant and the GCP are collaborating on integrating plant breeding information resources and applications into the infrastructure. This will automatically link these resources to upstream biological applications using the same infrastructure such as that used by the Systems Biology Knowledgebase initiative (http://genomicscience.energy.gov/compbio/#page=news) of the US Department of Energy which will be producing knowledge needed for crop improvement.

With all the progress achieved in marker technology, software development, analytical pipelines, and DMS, it is time to provide an IS, available through a public platform, that will offer breeding programs in developed and developing countries access to modern breeding technologies, in an integrated and configurable way, to boost crop quality and productivity.

Case Study: GCP's Integrated Breeding Platform

To fill this gap in the public sector and in particular in the arena of research for development, the GCP has been coordinating the development of the IBP (www.generationcp.org/ibp) in collaboration with scientists from ARIs, CGIAR centers, and national research programs since mid-2009. In a first phase the IBP aims at serving the needs of a set of 14 pioneer "user cases" – MB projects for eight crops in 16 developing countries in Africa and Asia. Leading scientists of those user cases help in testing the prototypes developed for the different tools of the analytical pipeline and contribute to the monitoring and evaluation of the platform development. This ensures that IBP development is driven by real breeding needs and its interface is user-friendly.

Objective of the IBP

The overall objective of the IBP project is to provide access to modern breeding technologies, breeding material, and related information and services in a centralized and functional manner to improve plant breeding efficiency in developing countries and hence facilitate the adoption of MB approaches. The short-term objective of the project (the initial phase) is to establish – through a client-centered approach – a minimum set of tools, data management infrastructure, and services to meet the needs and enhance the efficiency of the 14 user cases.

To achieve the overall objective, GCP is developing and deploying a sustainable IBP as a one-stop shop for information, analytical tools, and related services to design, implement, and analyze MB experiments. This platform should enable breeding programs in the public and private sectors to accelerate variety development for developing countries using marker technologies – from simple gene or transgene introgression to gene pyramiding and complex MARS and GWS projects. Hence IBP aims at bringing cutting-edge breeding technologies to breeding programs that are too resource-restricted to invest in the requisite genotyping and data management infrastructure and capacity on their own.

The IBP Partnerships

The primary stakeholders of the platform are plant scientists – at this time specifically breeders leading the selected MB projects of the 14 pioneer user cases. These pioneer user cases are all recently initiated marker-assisted breeding projects with specific budgets, objectives, and work plans. The needs of the projects are defining the user requirements, and hence the design and development prioritization of the different elements of the platform. In selecting the user cases, crop diversity was a primary consideration, since the platform is supposed to address the needs of a broad variety of crops. The platform's reciprocal contribution to these breeding projects is in helping them overcome bottlenecks that would compromise final product delivery and in enhancing their overall efficiency and chances of success by providing appropriate tools and support.

The developmental phase of the IBP brings together highly regarded public research teams – institutes and individuals who have been working on the challenges of crop information management and analysis, biometrics, and quantitative genetics. This team of bioinformaticians, statisticians, and developers aims to design and develop the different elements of the platform, based on needs and priorities defined by the user cases.

A continuous dialogue between users, developers, and service providers ensures a healthy balance between having a user-driven platform on the one hand, with a reasonable degree of "technology push" on the other hand, to ensure that users are kept abreast of technological solutions they may not be aware of but that would facilitate and accelerate breeding work.

The private sector has led the application of MB approaches and utilization of MBPs. The IBP is the first public sector effort of this magnitude aimed at developing and deploying an MBP. Given that MB for complex polygenic traits, and more so MARS, is still in its infancy in the public sector, it is recognized that efficient partnerships with the major private sector transnational seed companies is a strong prerequisite for the success of the IBP project. Consultations are ongoing with leaders in MB at Limagrain, Monsanto, Pioneer-DuPont, and Syngenta. Partnership with the private sector includes mainly some technology transfer, especially for stand-alone tools, and access to human resources to advise on the development of the platform and contribute to developing new tools or implement data management. The users, tools and services, and partnership of the platform are presented in Fig. 2.

The Platform

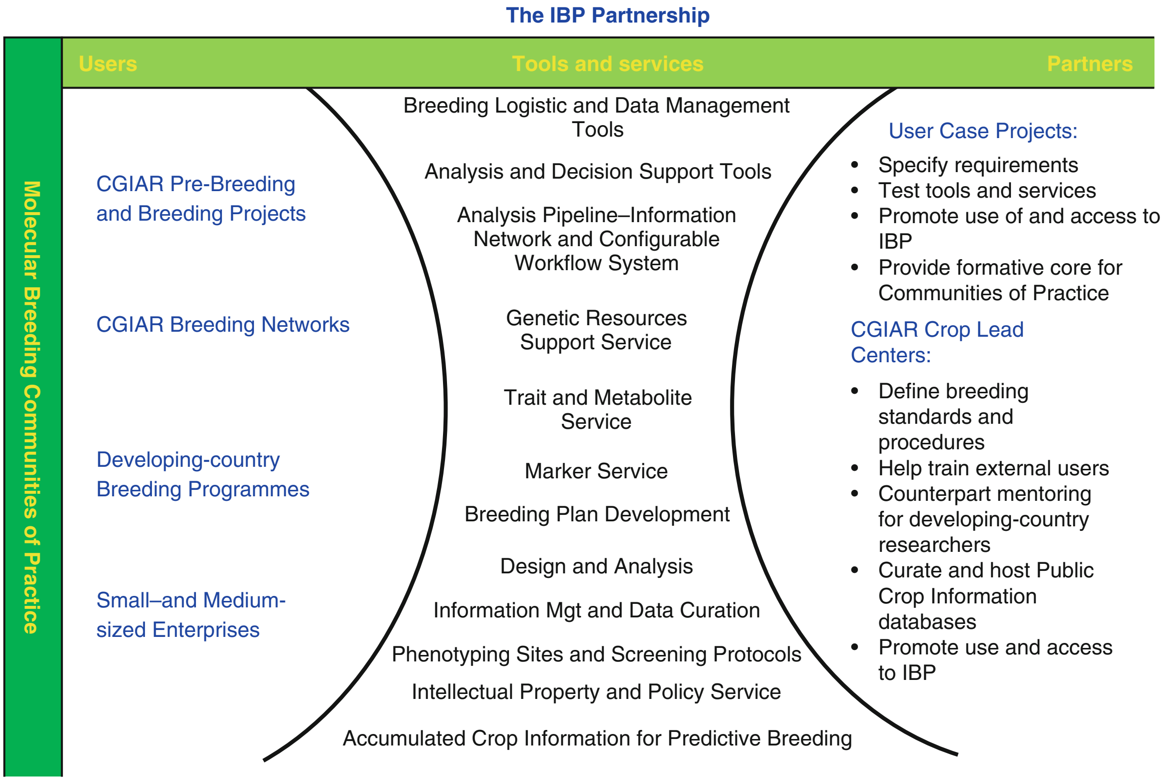
The IBP has three broad components (see Fig. 3): a Web-based portal and helpdesk, an open-source IS incorporating an adaptable breeding workflow system, and breeding and support services.

The stepwise development of the breeding workflow includes: (1) access to existing tools, (2) development of stand-alone new tools or adapted versions of existing tools to address the needs of the user cases, and (3) the integration of those tools into a CI (collaboration with the iPlant initiative) or through a thin middleware linking with local database to form a user-friendly configurable workflow system (CWS). A first version of the CWS, including an adequate set of tools, should be available by mid-2012, with full unfettered access scheduled for 2014.

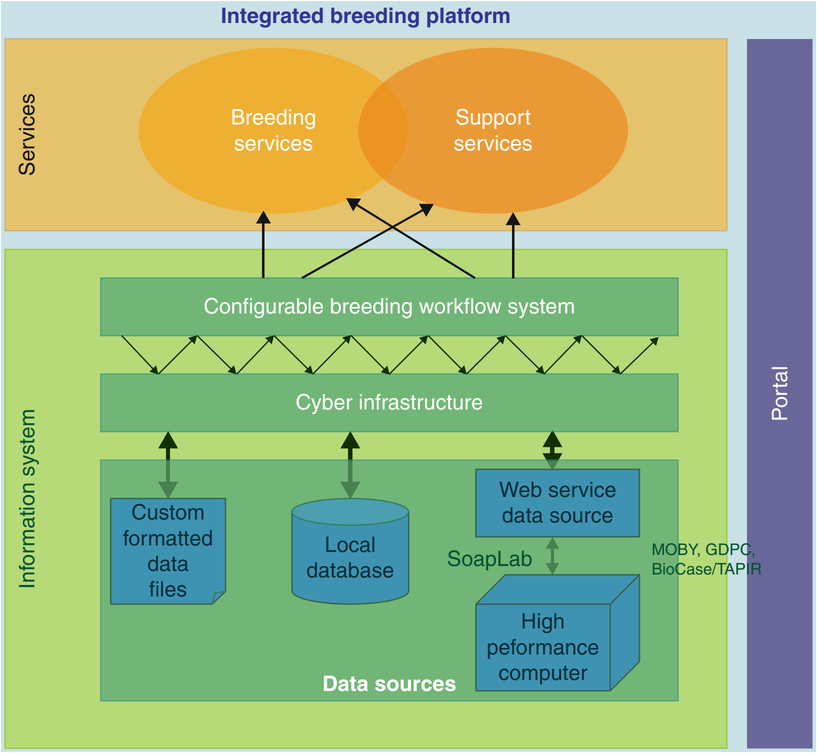
Component 1: The Integrated Breeding Portal and Helpdesk

Inaugurated by mid-2011, the portal is the online gateway through which users access all the tools and services of the IBP. Through the portal, users will select and download tools and instructions, order materials, and procure laboratory services.

The portal's helpdesk facilitates its use and ensures access for users who cannot efficiently use the Web interface by providing the elements they need via email, compact disc, and other offline media.


The IBP Partnership
Molecular Breeding Communities of Practice
Users
Tools and services
Partners
CGIAR Pre-Breeding and Breeding Projects
CGIAR Breeding Networks
Developing-country Breeding Programmes
Small–and Medium-sized Enterprises
Breeding Logistic and Data Management Tools
Analysis and Decision Support Tools
Analysis Pipeline–Information Network and Configurable Workflow System
Genetic Resources Support Service
Trait and Metabolite Service
Marker Service
Breeding Plan Development
Design and Analysis
Information Mgt and Data Curation
Phenotyping Sites and Screening Protocols
Intellectual Property and Policy Service
Accumulated Crop Information for Predictive Breeding
User Case Projects:
• Specify requirements
• Test tools and services
• Promote use of and access to IBP
• Provide formative core for Communities of Practice
CGIAR Crop Lead Centers:
• Define breeding standards and procedures
• Help train external users
• Counterpart mentoring for developing-country researchers
• Curate and host Public Crop Information databases
• Promote use and access to IBP

Molecular Breeding Platforms in World Agriculture. Figure 2
The IBP partnership



Molecular Breeding Platforms in World Agriculture. Figure 3
The IBP and its three main components

Through their user-friendly networking components, the Portal and Helpdesk will stimulate the development of collaborative crop-based and discipline-based communities of practice (CoPs). The CoPs are expected to promote the application of MB techniques and the utilization of facilitative information management technologies, enhance data and germplasm sharing, and generally advance modern breeding capacity by linking CGIAR Centers and ARIs with developing-country breeding programs and research organizations. There is a strong hope that CoPs will facilitate and accelerate a paradigm shift to a more collaborative, outward-looking, technology-enhanced approach to breeding.

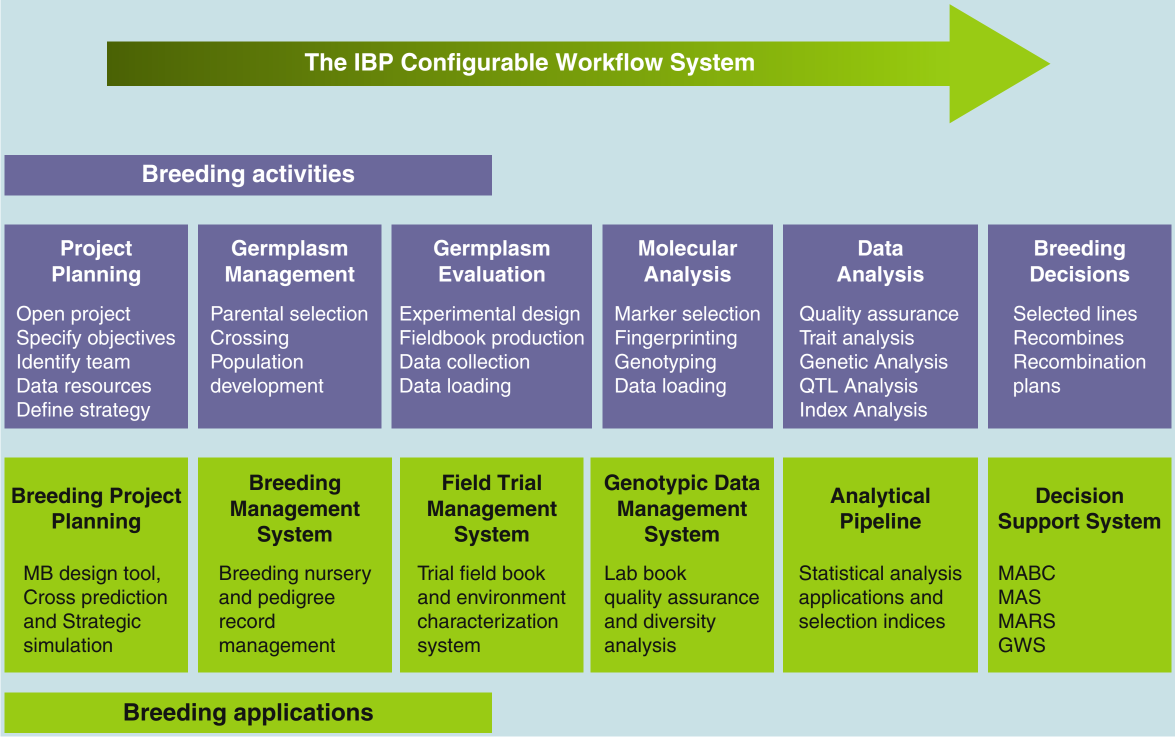
Component 2: The Information System

The IBP IS is structured as a CWS, with access to both local databases and distributed resources, such as central crop databases, molecular databases from GCP partner sites and from public initiatives such as Gramene and GrainGenes.

The Configurable Workflow System This CWS is the operational representation of the IS and will be implemented by assembling informatics tools into applications configured to match specific breeding workflows (e.g., for MAS, MABC, or MARS; Fig. 4). The tools are organized in a series of functional modules comprising the Integrated Breeding Workbench, which is really the background structure that implements the CWS.

The IBP CWS drives the users through the different practical steps or activities of an MB project. The setup of the experiment and the germplasm management are the first steps of any project, to be followed by a set of activities that can be repeated during subsequent crop cycles, depending on the breeding objective of the experiment:

- Germplasm evaluation
- Genetic analysis
- Data management
- Data analysis, and
- Breeding decisions


The IBP Configurable Workflow System

Breeding activities

Project Planning
Open project
Specify objectives
Identify team
Data resources
Define strategy

Germplasm Management
Parental selection
Crossing
Population development

Germplasm Evaluation
Experimental design
Fieldbook production
Data collection
Data loading

Molecular Analysis
Marker selection
Fingerprinting
Genotyping
Data loading

Data Analysis
Quality assurance
Trait analysis
Genetic Analysis
QTL Analysis
Index Analysis

Breeding Decisions
Selected lines
Recombines
Recombination plans

Breeding Project Planning
MB design tool, Cross prediction and Strategic simulation

Breeding Management System
Breeding nursery and pedigree record management

Field Trial Management System
Trial field book and environment characterization system

Genotypic Data Management System
Lab book quality assurance and diversity analysis

Analytical Pipeline
Statistical analysis applications and selection indices

Decision Support System
MABC
MAS
MARS
GWS

Breeding applications

Molecular Breeding Platforms in World Agriculture. Figure 4
The IBP configurable workflow system

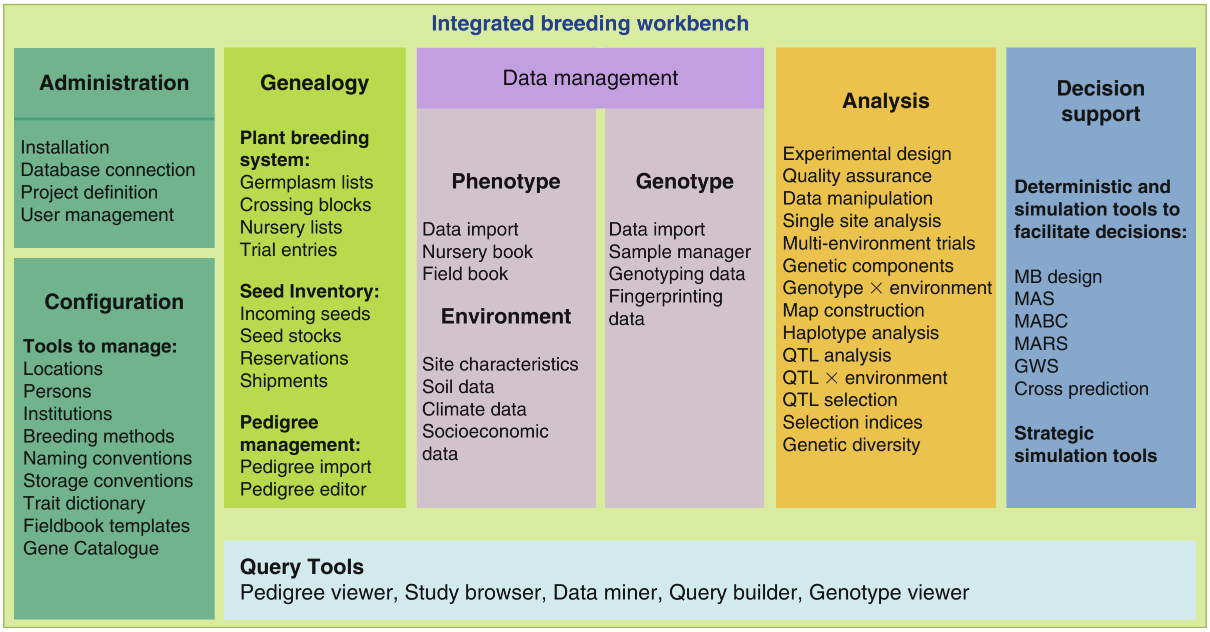
The Integrated Breeding Workbench The workbench starts as a blank slate and the first task for the user is to open or create a project. A project manages a breeding workflow for a particular crop and a specified user. The initial sets of tools which should be available are grouped in seven modules: Administration Tools, Configuration Tools, Query Tools, and Workflow Initialization Tools (genealogy, data management, analysis, and decision support; Fig. 5).

The administration module of the workbench specifies the crop, which identifies the central (public) data resources that will be accessible to the project. This includes a central genealogy database, a central phenotype database, a public gene management database, and a central genotype database. Each installation provides access to local (private) data resources. These data resources include a private or local database for the above data types as well as a seed inventory management system. Each installation has at least one user with administrative privileges. Users are identified by authentication codes (username and password) for access to specific private data resources. ("Private" simply means "requiring authentication for access" and several users may have access to the same private data.)

The first functionality of the workbench asks the user to open a project by selecting from a list of available project configuration "files." Once the configuration is selected, the availability of the public data resources should be checked, the user authentication codes verified, and the local data resources checked. Next, the list of modules should be reviewed and checked for availability and, depending on the state of the workflow, icons or menus should be made available for modules and tools.

The configuration tools allow users to:

- Select or specify naming conventions for germplasm, germplasm lists, studies, etc.
- Use and update ontologies such as germplasm methods and the trait dictionary
- Update breeding, testing, or collection locations
- Create and modify study templates



Integrated breeding workbench

Administration
Installation
Database connection
Project definition
User management

Configuration
Tools to manage:
Locations
Persons
Institutions
Breeding methods
Naming conventions
Storage conventions
Trait dictionary
Fieldbook templates
Gene Catalogue

Genealogy
Plant breeding system:
Germplasm lists
Crossing blocks
Nursery lists
Trial entries
Seed Inventory:
Incoming seeds
Seed stocks
Reservations
Shipments
Pedigree management:
Pedigree import
Pedigree editor

Data management

Phenotype
Data import
Nursery book
Field book

Environment
Site characteristics
Soil data
Climate data
Socioeconomic data

Genotype
Data import
Sample manager
Genotyping data
Fingerprinting data

Analysis
Experimental design
Quality assurance
Data manipulation
Single site analysis
Multi-environment trials
Genetic components
Genotype × environment
Map construction
Haplotype analysis
QTL analysis
QTL × environment
QTL selection
Selection indices
Genetic diversity

Decision support
Deterministic and simulation tools to facilitate decisions:
MB design
MAS
MABC
MARS
GWS
Cross prediction
Strategic simulation tools

Query Tools
Pedigree viewer, Study browser, Data miner, Query builder, Genotype viewer

Molecular Breeding Platforms in World Agriculture. Figure 5
The integrated breeding workbench

The query tools will depend on the data resources specified in the project configuration, and examples are:

- A germplasm and pedigree viewer
- A study browser to view phenotype or genotype data
- A data miner for identifying data patterns
- A cross-study query builder for linking different data sets
- A gene catalog viewer for viewing genetic diversity
- A genotype and trait viewer for visualizing graphical genotypes and trait markers

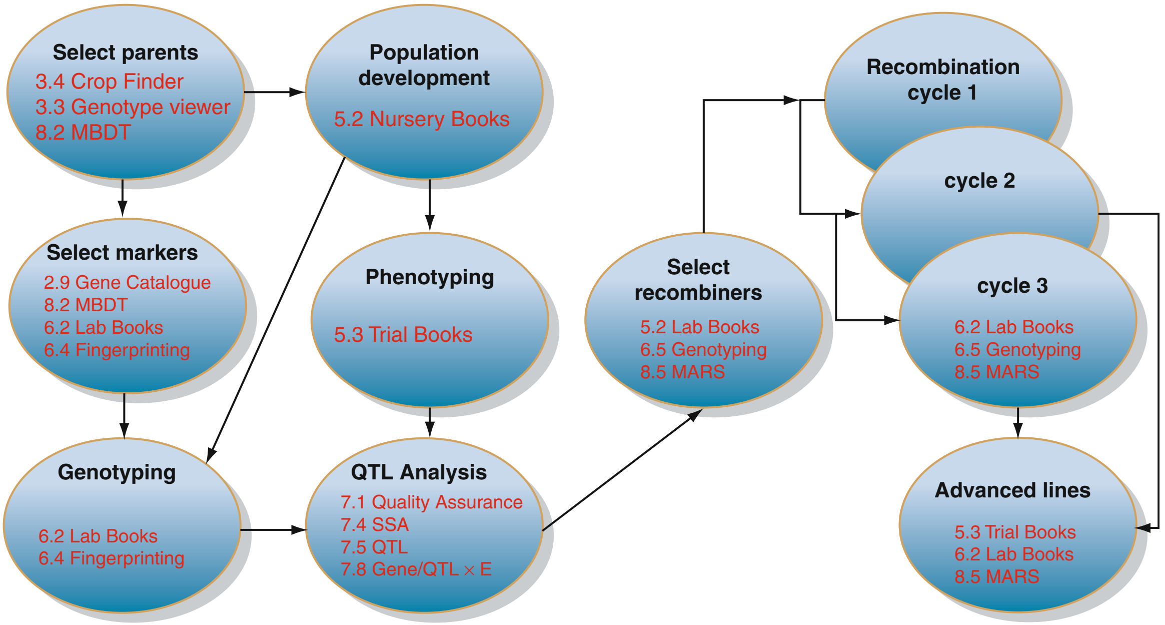
The workflow initialization tools comprise a set of modules (genealogy, data management, analysis, and decision support tools) that provide the user with a choice of different tools to achieve precise breeding objectives. Users might construct different breeding workflows to match their project activities. The user will only see the workbench tools and settings for those tools required to execute the steps in a particular breeding workflow, and at the appropriate step in that workflow.

The development of each tool is overseen by a team of IBP researchers, developers, and users who design, mock up, and prototype the tools of the breeding application and pass the specifications to a software engineering team. They will then monitor the development and test and support the application. For each application, the team develops a description of the application, functional specifications of all the tools, workflow specifications for the application, and an interface mockup. A workflow for a MARS project is shown in Fig. 6.

Component 3: IBP Services

The Services component comprises two modules. The first module, Breeding Services, provides services to conduct MB projects. The second module, Support Services, deals with training and capacity-building, aiming to provide support and improve capacity of NARS breeders to deliver improved germplasm through marker approaches – essential for the adoption of MB approaches and the MBP.

Breeding Services These services provide access to specific germplasm, and assist with contracting a service laboratory to conduct the marker work or to



Molecular Breeding Platforms in World Agriculture. Figure 6
Breeding workflow for an MARS experiment

quantify specific traits, such as metabolite profiles or grain quality parameters. The module has three elements (Fig. 7):

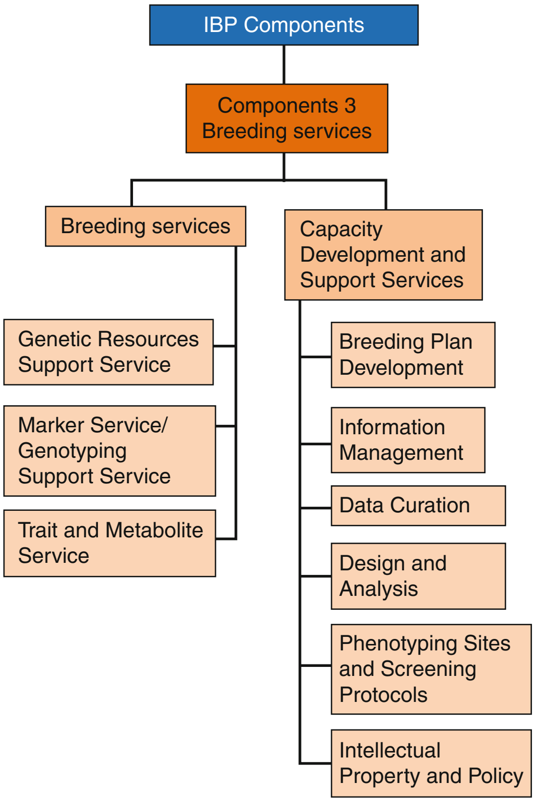
Genetic Resource Support Service: Access to suitable germplasm and related information from the different partners is a critical element of the portal. To address this, a Genetic Resource Support Service (GRSS) plans to tap into the CGIAR System-wide Genetic Resources Program (SGRP), a collaborative effort between GCP and existing gene banks in the CGIAR and NARS. The GRSS should ensure quality control, maintenance, and distribution of genetic resources, including reference sets and segregating populations acquired or generated through projects supported by GCP, and material generated from other sources and deposited with the GRSS (e.g., maize introgression lines from Syngenta).

Marker Service: The portal provides a set of online options for users to access different high-throughput marker service laboratories in the public and private sectors with clear contractual conditions. Service Laboratories have been selected on the basis of competitive cost, compliance with quality control requirements, and expeditious delivery, but are currently accessible by offline processes pending deployment of the IBP portal.

Trait and Metabolite Service: The portal provides a set of options for users to access laboratories specialized in the evaluation and analysis of specific traits, such as quality traits, pathology screening, or metabolite quantification. Analyses of certain secondary traits and metabolites that are indicative of plant stress tolerance can potentially provide valuable information to be used in breeding. Such analyses are generally prohibitively expensive if done locally, as it is difficult to maintain assay quality and devote the necessary resources for expertise, quality control, and specialized facilities.

Capacity Development and Support Services Capacity development is an integral part of the project, encompassing training and support in using MB techniques and markers, designing breeding strategies, quality data management, information analysis and decision modeling, phenotyping protocols, and protection of intellectual property (IP).

The main objective of this set of services is therefore to provide backstopping and training in a broad set of disciplines, to complement the elements of the breeding services and address specific technical and logistical bottlenecks. Such expert assistance is essential for the


IBP Components
Components 3
Breeding services
Breeding services
Capacity Development and Support Services
Genetic Resources Support Service
Marker Service/ Genotyping Support Service
Trait and Metabolite Service
Breeding Plan Development
Information Management
Data Curation
Design and Analysis
Phenotyping Sites and Screening Protocols
Intellectual Property and Policy

Molecular Breeding Platforms in World Agriculture. Figure 7
Organogram of the services provided by the IBP

adoption and proper use of new technologies. Services that will be available include:

Breeding plan development: It is essential to develop a breeding plan with a cost–benefit analysis before conducting a multi-cycle MB project. Depending on the nature of the experiment, such a plan may be quite simple or very elaborate, from the transfer of a single region (e.g., transgene) to complex selection that can consider the simultaneous transfer of dozens of regions. The critical factor is that the plan must detail all the activities over time, and the costs and benefits of the project to determine if it is worthwhile conducting the experiment. The platform provides templates and associated cost calculation sheets for different breeding schemes.

Information management: Under this service, assistance is provided in installing and parameterizing the platform IS for use by specific breeding projects.

Data curation: This service assists with capturing and curating current data for particular breeding projects, and in entering them into the integrated IS. This step is absolutely critical for quality control and further sharing of the information, and a contact person for each of the pioneer user cases has been identified to ensure good communication between the platform and the users.

Design and analysis: This service provides support on statistics, bioinformatics, quantitative genetics, and molecular biology. It includes training in data generation, handling, processing, and interpretation, as well as experimental design from field planting to MAS and MABC schemes. It provides assistance with the "translation" of the molecular context to the breeding context, and it will ensure that the methodology developed for the design and analysis of breeding trials is rapidly available to the users.

Phenotyping sites and screening protocols: Through this service, users can access information on phenotyping sites, protocols, and potential collaborators to ensure that selection is carried out under appropriate biotic and abiotic stresses and that the adaptation of germplasm is well characterized. Characterization of phenotypic sites includes geographical information, meteorological historical data, soil composition, and field infrastructure.

Genotyping Support Service (GSS): The GSS aims to facilitate access by developing country national agricultural research institutes to genotyping technologies, and bridge the gap between lab and field research. This service provides financial and technical support for NARS breeders to access cost-efficient genotyping services worldwide and supports training activities in experimental design and data analysis for MB projects.

Intellectual property (IP) and policy: This service provides support on IP rights and freedom to operate in the arena of biotechnology and germplasm use. The service is currently being provided on an experimental basis through a virtual IP Helpdesk hosted by the GCP web site at http://www.generationcp.org/iphelpdesk.php.

Integrated Breeding Hubs

If today few question the usefulness of local basic laboratories, it is also generally accepted that large-scale genotyping activities are best outsourced to cost-effective, high-throughput service laboratories, irrespective of location. Following that rationale, the IBP provides access to marker service laboratories as the main avenue to generate the large amount of genotyping data that will be necessary to support the extensive MABC programs of the future, starting with the user cases, but the GCP also recognizes the need to provide breeders in developing countries with access to some regional hubs. At the beginning of the project four regional hubs are envisioned, covering the needs of the Americas – Centro Internacional de Agricultura Tropical (CIAT, www.ciat.cigiar.org); Africa – BioSciences eastern and central Africa (BecA, http://hub.africabiosciences.org); South Asia – International Crops Research Institute for the Semi-Arid Tropics (ICRISAT, www.icrisat.org); and South East Asia – International Rice Research Institute (IRRI, www.irri.org).

These regional hubs are expected to provide the following services:

- In-house hands-on training (different formats are possible from short- to medium-length periods), with the objective of exposing scientists to new technologies and their applications to breeding.
- Training courses for selected groups of researchers, targeting basic knowledge of marker technologies and their applications, as well as data analysis. These courses can be used for the testing and validation of learning materials, which will then be continuously upgraded.
- Facilitation of small genomic and genotyping projects led by national programs, academia, and small and medium enterprises (SMEs).
- Marker services for "small" and "orphan" crops that do not have mass demand from breeding programs and would therefore not benefit from large service providers, due to the lack of availability of SNP markers and the need to use lower-throughput SSR or other markers that can more easily be handled in lower-tech laboratories.

The Genomics and Molecular Breeding Hubs should help raise the visibility of the IBP and thus help promote the adoption of MB. Collaboration between the IBP and the regional hubs is anticipated to occur through sharing information, guiding users to apply for the appropriate service, organizing training events, and planning other developments of common interest.

Scope and Potential for Molecular Breeding Platforms

Gaps Across Countries and Crops

The application of MB approaches is now routine in developed countries, as is the integration of facilitative information and communication technologies, which are critical given the immense volumes of data necessary for, and generated by, these breeding processes. However, the situation is very different in developing countries, where MB is still far from routine in its application in breeding programs, particularly in Africa. This is especially critical due to the monumental and urgent imperative to rapidly achieve food security and improve livelihoods for a rapidly growing population through breeding for biotic stresses (including weeds, pests, and diseases) and abiotic stresses (including physical soil degradation, nitrogen deficiency, drought, heat, cold, and salinity) – conditions that make accurate phenotyping challenging. Fortunately, the history of modern breeding in developing countries is comparatively short, allowing a larger potential for crop improvement relative to the genetic gains that can be obtained at this time in developed countries, in which extensive breeding has been applied to crops for a longer time.

To address these issues, the capacity of national research institutions in terms of funds, infrastructure and expertise is directly related to the strength of their national economies [86]. This is reflected in the sharp differences in the capacity to conduct and apply biotechnology research as observed across developing countries (FAOBioDeC, http://www.fao.org/biotech/inventory_admin/dep/default.asp), and by the same token in their capacity to establish and/or utilize MBPs. The result is a three-tier typology of developing countries, directly attributable to the level of each country's investment in agricultural R&D [87].

Tier-1 countries, comprising newly industrialized countries (NICs) such as Brazil, China, India, Mexico, South Africa, and Thailand, substantially invest in technology and R&D and are self-reliant in most aspects of marker technologies [88, 89]. These countries have the simultaneous potential to effectively adopt, adapt, and apply information and communication technologies to enhance research efficiency and outputs. They are therefore naturally at the vanguard in adopting MBPs.

Mid-level developing world economies (tier-2) such as Colombia, Indonesia, Kenya, Morocco, Uruguay, and Vietnam are well aware of MB's importance, and some effectively apply marker technologies for germplasm characterization [90–93] and selection of major genes [94–99]. These countries have a matching potential for a limited utilization of MBPs, a potential that can be enhanced fairly rapidly in the medium to long term.

Low-level developing world economies (tier-3 countries) are struggling to sustain even basic conventional breeding. They have very limited or no application of MB approaches and are unlikely to adopt MBPs except in the long term.

Especially for tier-3 countries, resource-limited breeding programs in many developing countries are severely hampered by a shortage of well-trained personnel, low level of research funding, inadequate access to high-throughput genotyping capacity, poor and inadequate phenotyping infrastructure, lack of ISs and appropriate analysis tools, and by the logistical difficulty of integrating new approaches with traditional breeding methodologies – including problems of scale when scaling up from small to large breeding programs.

Until recently, the scarcity of available genomic resources for clonally propagated crops, for some neglected cereals such as millet, and for less-studied crops such as most tropical legumes, which are all very important crops in developing countries, represented a further constraint to agricultural research for development [100], thereby limiting the application of molecular approaches and hence the potential for MBPs. However, the recent emergence of affordable large-scale marker technologies (e.g., DArT [101]), the sharp decline of sequencing costs boosting marker development based on sequence information [102], and the explicit efforts of national agricultural research programs (e.g., India [103]) and international initiatives such as GCP [104]) have all resulted in a significant increase in the number of genomic resources available for less-studied crops. As a result, most key crops in developing countries now have adequate genomic resources for meaningful genetic studies and most MB applications.

Similarly, international efforts such as GCP's IBP are designed to help overcome the challenges of developing-country breeders – exploiting economies of scale by making available convenient and cost-effective collective access to cutting-edge breeding technologies and informatics hitherto unavailable to them, including genomic resources, advanced laboratory services, and robust analytical and data management tools. Together, this increasing availability of genomic resources and tools for previously neglected but important crops and the access to initiatives targeting the resource-challenged NARS of the developing world will hasten the adoption of MBPs for these countries.

Institutional, Governmental, and Public Support

While corporate and other proprietary MBPs need only meet the specific requirements of a particular corporation or of specific paying clients, the development of platforms targeted at breeding programs in the developing world require a broad consensus among the parties that would use them and support them from multiple overseeing organizations. This is because these platforms are built on the premise of minimizing costs and maximizing benefits through economies of scale generated through collective access by multiple partners.

The public-access MBPs would therefore be critically dependent on well-structured MB programs, which may not be a reality in many developing countries. A good structure would entail compliance with common or compatible:

- Good field infrastructure, including meteo station
- Good agronomical practices at experimental stations
- Crop ontology information system
- Data collection, management, and analysis protocols
- Breeding plan design
- Information and communication technology infrastructure
- Informatics tools for analysis, decision support purposes, and eventually modeling and simulation

Traditionally, developing world breeding programs have largely been poorly funded and poorly supported, and have been primarily driven by donor organizations [105, 106]. The lack of in-country support has often limited the dependent breeding activities to no more than a basic level. Under such circumstances, it was unrealistic to anticipate the adoption of new biotechnologies – including the utilization of MBPs. Fortunately, this scenario is changing. In 2003, through the Comprehensive Africa Agriculture Development Programme (CAADP, http://www.caadp.net/implementingcaadp-agenda.php), African governments committed to invest more in food security and in agriculture-led growth. Since then, many countries in Africa and elsewhere have developed comprehensive agricultural development strategies.

There is also a growing participation by foundations and nongovernmental organizations, and more recently the emergence of public–private sector partnerships (e.g., US Global Food Security Plan, http://www.state.gov/s/globalfoodsecurity/129952.htm). This governmental and institutional commitment is critical for the adoption of biotechnologies in general [8, 107] and for MB adoption in tier-2 countries in particular, with the attendant establishment and utilization of MBPs.

Challenges, Risks, and Opportunities

Challenges hampering the potential of MBPs in developing countries include both factors applicable generally to MB and those specific to MBPs. These factors encompass infrastructure capacity, human resource, and operational and policy issues. But amidst the challenges there are also actual and potential opportunities.

Human Capacity Human capacity for MB technologies in developing countries is a challenge, and limitations include substandard agriculture programs at universities; difficulties in keeping up to date with relevant developments, including failures by others; poor technical skills in core disciplines; isolation as a result of insufficient peer critical mass in the workplace; and poor incentives to attract and retain scientists, resulting in brain drain and staff turnover [108].

To partially offset the undesirable trend of losing the "champions" and to "generate" more "champions," novel international initiatives like Alliance for a Green Revolution in Africa (AGRA) support high-quality education in the South. Examples include the African Centre for Crop Improvement (ACCI, http://www.acci.org.za/) based at the University of KwaZulu–Natal in South Africa and the University of Ghana-based West African Centre for Crop Improvement (WACCI, http://www.wacci.edu.gh/). Both institutes offer doctorate degrees in modern breeding to African students, with the fieldwork component being carried out in the students' home countries.

While obtaining their Ph.D. in plant breeding, these scientists study the principles of marker technologies, equipping them to undertake MB activities. To retain this much-needed expertise in Africa, the WACCI and ACCI programs also provide post-Ph.D. funds for these scientists to conduct research in their home countries and, in some cases, provide matching funds for their career advancement.

Precise Phenotyping There can be no successful MB program without precise phenotyping of the target traits. Reliable phenotypic data is a must for good genetic studies [109] and most developing countries lack suitable field infrastructure for good trials and collection of accurate phenotypic data. As part of the services of a good MBP, guidelines on best practice must be provided on how to design and run a trial and conduct precise phenotyping for genetic studies under different target environments. Improving access to homogeneous field areas, and paying attention to good soil preparation and homogeneous sowing are critical. The development of new geographic IS tools [102, 110], experimental designs, phenotyping methodologies [111, 112], and advanced statistical methods [113] will facilitate the understanding of the genetic basis of complex traits [114] and of genotype-by-environment (G×E) interactions [48, 115]. Improving phenotyping infrastructure in developing countries must thus be a top priority to promote modern breeding and utilization of MBPs [106].

Laboratories for Markers Services Genotyping can be expensive when it is performed in small laboratories using labor-intensive and low-throughput markers such as SSRs. This has traditionally limited the use of MMs in developing countries beyond the fingerprinting of germplasm with a small number of markers or the use of MAS for a few key traits. Operational efficiency is also vital, because fundamental timelines must be respected to ensure that no crop cycle is lost. Indeed, at every selection cycle, a service laboratory may have only a few weeks (time between DNA being extracted from leaves harvested on plantlets and the flowering time) to conduct the analysis and return the data to the breeders to enable them to conduct appropriate crosses among selected genotypes.

There is general agreement today that basic local laboratories at national and regional levels can be useful at least to service small local needs such as fingerprinting of limited number of accessions, GMO detection or MAS for specific traits, or for teaching and training purposes. It is also generally accepted that large-scale genotyping activities are best outsourced to advanced, modern, cost-effective high-throughput service laboratories, irrespective of the original location of the needs. This outsourcing is driven by the evolution in marker technologies. The advent of SNP genotyping led the shift from the low-throughput, primarily manual world of SSRs to high-throughput platforms powered by robotics and automated scoring, better handled by dedicated service laboratories [102, 116, 117]. As a result, genotyping costs have decreased by up to tenfold while data throughput has increased by the same magnitude. An example for MARS is provided in Fig. 6. SNP markers are increasingly available for most mainstream crops and for several less-studied crops [118, 119], which are important in developing countries.

A particular effort will be needed to ensure an easy and reliable way to track samples from the field to the laboratory, and back to the field – it will hence be vital to carefully identify DNA samples from material collected in the field. Such documentation should optimally be through bar-coding, and all information pertaining to management of field trials or experiments should be recorded in electronic field books. Marker work would of necessity be subcontracted to a service lab with a good and preferably platform-compatible laboratory information management system (LIMS).

Data Management For breeders to efficiently access relevant information generated by themselves and by other researchers, reliable data management (including sample tracking, data collection and storage, and modern analytical methodologies and tools for accurate decision making, among others) is critical both within a given MB program and across programs. In view of this, it is essential that breeders manage pedigree, phenotypic, and genotypic information through common or mutually compatible crop databases, in keeping with the collective access principle of a public MBP. The format of databases would need to be user-friendly and compatible with field data collection devices and applications to encourage both adoption and compliance. Ultimately, data collection and management processes would need to seamlessly link with a platform-resident analysis, modeling, simulation, and a decision support workbench for full utility of the breeding platform.

Paradigm Shift: Collaborative Work and Data Sharing Access to information and products generated by fellow users is a potentially critical incentive for breeders to use the platform and share their own data with other users. However, this would require a fundamental paradigm shift from the present data-hoarding, inward-looking approach to research common to breeders. This may, however, only be achievable if it is a clear requirement in the terms of engagement for membership of a "platform community," or if distinct financial and other incentives are offered for such sharing.

Technology-Push Versus Demand-Driven An MBP is by nature a high-level technological solution. It carries with it the inherent risk of failing to address fundamental practical problems of developing-world breeding programs, which will often by nature be technology-deficient. Such platforms therefore face the challenge of ensuring that they meet targeted user objectives and address practical constraints.

However, with this challenge comes an opportunity to introduce advanced MB methodologies to developing world breeders, by encouraging change that will enable them to take advantage of the efficiencies and economies of scale offered by the MBP. This opportunity would be particularly reachable with bottom-up platform design and development that actively engages and involves the breeders – including elements of human resource capacity development and support in usage.

Adoption and Use by Breeders An MBP would only make a difference if it is adopted and widely used by the breeders. The most important element influencing this would be credibility – a function of the quality of the technology, the awareness of potential users, the ease of access, and initial incentives. There is a need for successful public sector developing-country examples to demonstrate that the platform can effectively enhance the efficiency of breeders through the use of modern approaches – a clear demonstration of the added value of using the platform.

Sustainability of the Platform Sustainability would be a challenge for MBPs targeting developing world breeding programs, given their resource limitations. These programs may not be able to meet the full cost of platform usage, and the cost of maintaining and updating the different elements of the platform on a regular basis – particularly tools and facilities that must keep abreast with evolving information and communication technologies.

Of course, platform sustainability is directly linked to its adoption by breeders, and sustainability strategies must be adapted to the diversity and financial resources of the potential clients, from developing-world national agricultural research institutes with limited resources to SMEs. Service costs might also be adjusted if clients are willing to share data and release germplasm through the platform.

Platform managers may also have to consider other innovative options like on-platform advertising by agriculture-related commercial enterprises. However, ongoing donor support would most likely still be required in the medium to long term.

Communities of Practice The development of platform-based MB communities of practice, to connect groups of crop researchers, mainly breeders, willing to share experiences and information on modern breeding methods, best field practices, and development of improved varieties, and to practice peer-to-peer mentoring, are an additional potential avenue for platform adoption and sustainability, besides providing means to quickly and efficiently resolve recurring breeding problems. Partnerships between developed and developing-country institutions, and between the private and public sectors, are also an opportunity for realizing the full potential of MB [87, 108].

Many other hurdles limit successful public sector utilization of MB opportunities [120, 121]. However, the potential of virtual MBPs made possible by the revolution in information and communication technologies provides opportunities to counter and overcome many of those shortcomings.

Potential Economic Impact of Molecular Breeding Platforms

By its nature, MB improves the efficiency of crop breeding – progressively increasing genetic gains by selecting and stacking favorable alleles at target loci. The utilization of MBPs accelerates and amplifies the advantages of MB by introducing significant efficiencies in resource and time usage. Predictive or designer breeding, which would be the ultimate result of information-rich MB, attainable through the use of MBPs by numerous different breeding programs that freely share data and germplasm, would particularly bring about these savings in resources and time.

However, a direct comparison of the cost-effectiveness of MB with phenotypic selection is not straightforward. Firstly, factors other than cost – such as trade-offs between time and money – play an important role in determining the selection method. Secondly, this choice is further complicated by the fact that the two methods are rarely mutually exclusive or direct substitutes for each other [122]. On the contrary, under most breeding schemes, they are in fact complementary. Where operating capital is not a limitation, MB maximizes the net present value, especially when strengthened through MBPs [123]. With the increasing ease of accessing marker service laboratories and the declining cost per marker data point, MB costs are shrinking, making it extremely attractive from a purely economic perspective.

However, once the technological hurdles are overcome, the ultimate impact of new technologies (such as MBPs) is often limited by the lack of, or ineffective, seed distribution systems or by distant markets. SMEs are critical in promoting access to, and distribution of, improved seeds, thus helping alleviate a major bottleneck to the impact of improved breeding on smallholder farmers [124, 125].

Few economic analyses have been conducted to objectively assess the potential impacts of MB in the public sector, and none for MBPs that are just now emerging as a tool for breeding in the public sector.

Of the few analyses done to date, one evaluates the economic benefits of MABC using preexisting MMs in developing rice varieties tolerant to salinity and P-deficiency [126] in Bangladesh, India, Indonesia, and the Philippines. Encompassing a broad set of economic parameters, the study concluded that MABC saves an estimated minimum of 2–3 years, resulting in significant incremental benefits in the range of USD 300–800 million depending on the country, the extent of abiotic stress encountered, and the lag for conventional breeding [127].

Future studies are likely to confirm the positive economic benefits of MB and, given that MBPs amplify the benefits of MB, it can be reasonably inferred that the emerging platforms would indeed further enhance those economic benefits.

Future Directions

MBPs will inevitably have a significant impact on crop breeding in developing countries in the medium to long term because of:

- The needs-driven demand for improved crop varieties to counter the global food crisis
- The exponential development of genomic resources
- The ever-declining cost of marker technologies
- The increasing occurrence of public–private partnerships, where the public sector can learn from private companies about best practices for integrating MB into their breeding programs
- The need for innovative solutions to the challenges of resource and operational limitations

The first challenge of MBPs will be to meet the immediate needs of the breeders in developing-country public and private programs. The first step will be to provide them with the tools for enhancement of their current breeding programs, through the implementation of field books, pedigree management, and basic statistical analytical tools necessary to optimally conduct their current breeding efforts. In close succession with these first applications, tools will need to be made available to facilitate the integration of MB into their breeding programs. Databases will need to be developed for storing genotypic and phenotypic data, integrated analytical tools will need to be made available to breeders for analysis of this accumulated data and for the identification of important simple trait loci or QTLs to monitor and recombine in their breeding programs, and decision support tools will need to be developed to help breeders decide on the next steps to engage in based on the data they generated from their MB activities.

In the near future, more complex tools will need to be developed for the storage and analysis of the large amounts of genotypic data that will be generated by new next-generation sequencing technologies and for their application in GWS. A tight linkage will also have to be established with the wealth of information that is being generated and will continue to be generated even faster in the genomics area, leading to the dissection of the genome and to the discovery of the location and function of major genes having an impact upon the performance of crops in environments relevant to developing-country programs.

Eventually, the accumulation of large amounts of genetic information linked to specific haplotypes will lead to the increasing use of predictive breeding in combination with traditional MB usage and appropriate tools will also need to be developed to support those efforts.

Although it is critical for a platform to anticipate all the new possible features of MB, ensuring that new technologies and ISs will find their way in a flexible infrastructure, it is also quite probable that most of the breeding programs in developing countries will work at the short- and mid-term mainly with simple MB approaches as they will never reach the critical size of crosses and germplasm evaluation requested to maximize complex approaches.

Conclusion and Prospective Scenarios

Through international initiatives like the ones coordinated by the CGIAR centers and programs, several notable developing-world MB successes have already been reported.

A well-known example is the development of submergence-tolerant rice cultivars through MABC led by IRRI [128]. The introgression of the Sub1 gene from

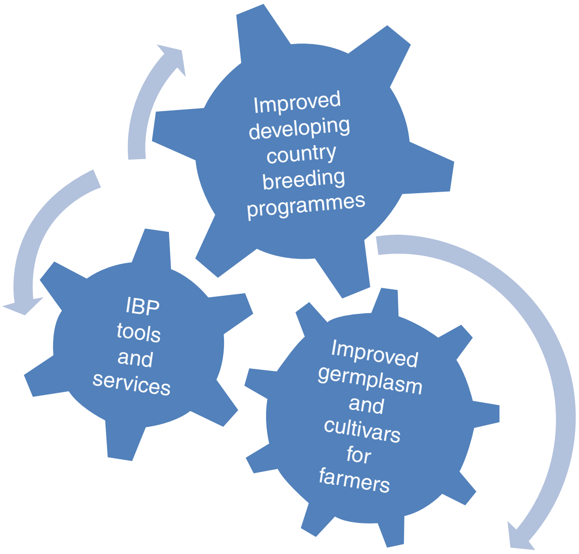
FR13A (the world's most flood-tolerant variety) into widely grown varieties like Swarna improved yields in more than 15 million hectares of rain-fed lowland rice in South and Southeast Asia.

MB in general and the use of MBPs in particular have definitely been shown to be an efficient approach for reducing the number of required selection cycles and for increasing the genetic gain per crop cycle to a point where the required human and operational resources can be kept to a minimum.

However, for sustainable adoption, the use of modern breeding strategies requires a breeder-led bottom-up approach. As a start, simple MB approaches adapted to local environments should be tested first by individual breeders to evaluate their success and impact under those breeders' conditions. Once proven, these approaches can then be implemented more widely or integrated to an MBP for enhanced efficiency. In case of individual success the adoption of MB by those breeders should be quite straightforward.

It is clear that the extent, speed, and scope of adoption of MB approaches and of utilization of MBPs will vary somewhat across tier-1, tier-2, and tier-3 countries, depending on the local priorities and on the resources available in given breeding programs. It is unrealistic to expect that large-scale MB breeding activities, including utilization of MBPs, will be widely implemented across the board in developing countries in the near term. However, the prospects are bright for individual breeders in these countries (particularly in tiers 1 and 2) to access germplasm, data, tools, and methodology that will allow them to conduct efficient MB projects by taking advantage of large international initiatives specifically targeting developing-country breeding programs. This will, however, happen in different ways and on different timelines for each tier.

For tier-1 countries, the impact would be evident in the shorter term – say in 3–6 years. These countries will benefit from new tools and platforms by increasing the rate of MB adoption. The biggest change is likely to occur in tier-2 countries, as these countries would be starting MB from scratch, but the impact would realistically be measurable only in the medium term, meaning in about a decade from now. For countries currently in tier-3 to advance to tier-2, basic breeding programs must first be established, which is highly dependent on governmental priorities and on subsequent resource allocation.


Improved developing country breeding programmes
IBP tools and services
Improved germplasm and cultivars for farmers

Molecular Breeding Platforms in World Agriculture. Figure 8
IBP as a key component to boost NARS breeding capacities and therefore crop productivity in developing countries

All in all, implementing MB (and catalyzing and accelerating its impact through MBPs) will boost crop production, which will translate into higher farm productivity per unit of land, better nutrition, higher incomes, poverty alleviation, and ultimately improved livelihoods in developing countries (Fig. 8). These gains will be amplified by sustained use, by continuously improving expertise, and by growth and development of homegrown capacity for the application of advanced breeding approaches.

Bibliography

1. Crosbie TM, Eathington SR, Johnson GR, Edwards M, Reiter R, Stark S, Mohanty RG, Oyervides M, Buehler RE, Walker AK, Dobert R, Delannay X, Pershing JC, Hall MA, Lamkey KR (2006) Plant breeding: past, present, and future. In: Lamkey KR, Lee M (eds) Plant breeding: the Arnel R. Hallauer international symposium. Blackwell, Ames, pp 3–50
2. Falck-Zepeda J, Zambrano P, Cohen JI, Borges O, Guimarães EP, Hautea D, Kengue J, Songa J (2008) Plant genetic resources for agriculture, plant breeding, and biotechnology. EPTD Discussion Paper 00762. International Food Policy Research Institute, Washington, DC

3. Goodman RM, Hauptli H, Crossway A, Knauf VC (1987) Gene transfer in crop improvement. Science 236:48–54
4. Cooper M, Smith OS, Merrill RE, Arthur L, Polich DW, Loffler CM (2006) Integrating breeding tools to generate information for efficient breeding: past, present, and future. In: Lamkey KR, Lee MA (eds) Plant breeding: the Arnel R. Hallauer international symposium. Blackwell, Ames, pp 141–154
5. Tanksley SD, Young ND, Paterson AH, Bonierbale MW (1989) RFLP mapping in plant breeding: new tools for an old science. Biotechnology 7:257–264
6. Ribaut J-M, Hoisington DA (1998) Marker-assisted selection: new tools and strategies. Trends Plant Sci 3:236–239
7. Bernardo R (2008) Molecular markers and selection for complex traits in plants: Learning from the last 20 years. Crop Sci 48:1649–1664
8. Moose SP, Mumm RH (2008) Molecular plant breeding as the foundation for 21st century crop improvement. Plant Phys 147:969–977
9. Wang S, Basten CJ, Zeng Z-B (2005) Windows QTL Cartographer 2.5. Department of Statistics, North Carolina State University, Raleigh
10. Collard BCY, Mackill DJ (2008) Marker-assisted selection: an approach for precision plant breeding in the twenty-first century. Philos Trans R Soc B 363:557–572
11. Ribaut J-M, Jiang C, Hoisington D (2002) Efficiency of a gene introgression experiment by backcrossing. Crop Sci 42:557–565
12. Mumm RH (2007) Backcross versus forward breeding in the development of transgenic maize hybrids: theory and practice. Crop Sci 47(S3):S164–S171
13. Hospital F, Charcosset A (1997) Marker-assisted introgression of quantitative trait loci. Genetics 147:1469–1485
14. Stam P (1995) Marker-assisted breeding. In: Van Ooijen JW, Jansen J (eds) Biometrics in plant breeding: applications of molecular markers. Proceedings of the ninth meeting of the EUCARPIA section biometrics in plant breeding, CPRO-DLO, Wageningen, pp 32–44
15. Peleman JD, Van Der Voort JR (2003) Breeding by design. Trends Plant Sci 7:330–334
16. Johnson R (2004) Marker-assisted selection. Plant Breed Rev 24:293–309
17. Bernardo R, Charcosset A (2006) Usefulness of gene information in marker-assisted recurrent selection: a simulation appraisal. Crop Sci 46:614–662
18. Guttmacher AE, Collins FS (2002) Genomic medicine – a primer. N Engl J Med 347:1512–1520
19. de los Campos G, Gianola D, Allison DB (2010) Predicting genetic predisposition in humans: the promise of whole-genome markers. Nat Rev Genet 11:880–886. doi:10.1038/nrg2898
20. Goddard ME, Hayes BJ (2007) Genomic selection. J Anim Breed Genet 124:323–330
21. Tinker NA, Yan W (2006) Information systems for crop performance data. Can J Plant Sci 86:647–662
22. Yan W, Tinker NA (2007) DUDE: a user-friendly crop information system. Agron J 99:1029–1033
23. McLaren CG, Bruskiewich RM, Portugal AM, Cosico B (2005) The international rice information system. A platform for meta-analysis of rice crop data. Plant Physiol 139:637–642
24. Bruskiewich R, Senger M, Davenport G, Ruiz M, Rouard M, Hazekamp T, Takeya M, Doi K, Satoh K, Costa M, Simon R, Balaji J, Akintunde A, Mauleon R, Wanchana S, Shah T, Anacleto M, Portugal A, Ulat VJ, Thongjuea S, Braak K, Ritter S, Dereeper A, Skofic M, Rojas E, Martins N, Pappas G, Alamban R, Almodiel R, Barboza LH, Detras J, Manansala K, Mendoza MJ, Morales J, Peralta B, Valerio R, Zhang Y, Gregorio S, Hermocilla J, Echavez M, Yap JM, Farmer SA, Gary, Lee J, Casstevens T, Jaiswal P, Meintjes A, Wilkinson M, Good B, Wagner J, Morris J, Marshall D, Collins A, Kikuchi S, Metz T, McLaren G, van Hintum T (2008) The Generation Challenge Programme platform: semantic standards and workbench for crop science. J Plant Genom 2008, Article ID 369601, 6 p. doi: 10.1155/2008/369601
25. Rodgers D, Jordan D (2009) Information management systems for plant breeders. Primary Industries and Fisheries (PI&F) of the Queensland Government, Department of Employment, Economic Development and Innovation in Australia, Queensland, Australia
26. Gudmundur A, Thorisson JM, Brookes AJ (2009) Genotype–phenotype databases: challenges and solutions for the post-genomic era. Nat Rev 10:9–18
27. Smith A, Cullis B, Thompson R (2001) Analyzing variety by environment data using multiplicative mixed models and adjustments for spatial field trend. Biometrics 57: 1138–1147
28. Burgueño J, Crossa J, Cornelius PL, Trethowan R, McLaren G, Krishnamachari A (2007) Modeling additive × environment and additive × additive × environment using genetic covariances of relatives of wheat genotypes. Crop Sci 47:311–320
29. Butler D, Cullis BR, Gilmour AR, Gogel BJ (2007) ASReml reference manual, release 2.00. VSN, Hemel Hempstead
30. Hammer G, Cooper M, Tardieu F, Welch S, Walsh B, van Eeuwijk F, Chapman S, Podlich D (2006) Models for navigating biological complexity in breeding improved crop plants. Trends Plant Sci 11:587–593
31. Wang J, Chapman SC, Bonnett DG, Rebetzke GJ, Crouch J (2007) Application of population genetic theory and simulation models to efficiently pyramid multiple genes via marker-assisted selection. Crop Sci 47:580–588
32. Chapman S (2008) Use of crop models to understand genotype by environment interactions for drought in real-world and simulated plant breeding trials. Euphytica 161:195–208
33. DeLacy IH, Fox PN, McLaren G, Trethowan R, White JW (2009) A conceptual model for describing processes of crop improvement in database structures. Crop Sci 49:2100–2112
34. Crossa J, Burgueño J, Dreisigacker S, Vargas M, Herrera S, Lillemo M, Singh RP, Trethowan R, Franco J, Warburton M, Reynolds M, Crouch JH, Ortiz R (2007) Association analysis of historical bread wheat germplasm using additive genetic covariance of relatives and population structure. Genetics 177:1889–1913

35. Lee M (1995) DNA markers and plant breeding programs. Adv Agron 55:265–344
36. Helentjaris T, Slocum M, Wright S, Schaefer A, Nienhuis J (1986) Construction of genetic linkage maps in maize and tomato using restriction fragment length polymorphisms. Theor Appl Genet 72:761–769
37. Edwards MD, Stuber CW, Wendel JF (1987) Molecular-marker-facilitated investigations of quantitative-trait loci in maize. I. Numbers, genomic distribution and types of gene action. Genetics 116:113–125
38. Paterson AH, Lander ES, Hewitt JD, Peterson S, Lincoln SE, Tanksley SD (1988) Resolution of quantitative traits into Mendelian factors by using a complete linkage map of restriction fragment length polymorphisms. Nature 335:721–726
39. Mullis K (1990) The unusual origin of the polymerase chain reaction. Sci Am 262:56–65
40. Senior ML, Heun M (1993) Mapping maize microsatellites and polymerase chain reaction confirmation of the target repeats using a CT primer. Genome 36:884–889
41. Vos P, Hogers R, Bleeker M, Reijans M, Tho L, van der Hornes M, Frijters A, Pot J, Peleman J, Kuiper M, Zabeau M (1995) AFLP: a new technique for DNA fingerprinting. Nucleic Acids Res 23:4407–4414
42. Gilles PN, Wu DJ, Foster CB, Dillon PJ, Chanock SJ (1999) Single nucleotide polymorphic discrimination by an electronic dot blot assay on semiconductor microchips. Nat Biotechnol 17:365–370
43. Eathington SR, Crosbie TM, Edwards MD, Reiter RS, Bull JK (2007) Molecular markers in commercial breeding. Crop Sci 47:154–163
44. Lander ES, Botstein D (1989) Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics 121:185–199
45. Borevitz J (2004) Genomic approaches to identifying quantitative trait loci: lessons from *Arabidopsis thaliana*. In: Cronk QCB, Whitton J, Ree RH, Taylor IEP (eds) Molecular genetics and ecology of plant adaptation. Proceedings of an international workshop, December 2002, Vancouver, NCR Research Press, Ottawa, pp 53–60
46. Haley CS, Knott SA (1992) A simple regression method for mapping quantitative loci in line crosses using flanking markers. Heredity 69:315–324
47. Martinez O, Curnow RN (1992) Estimating the locations and the sizes of the effects of quantitative trait loci using flanking markers. Theor Appl Genet 85:480–488
48. Malosetti M, Ribaut J-M, Vargas M, Crossa J, van Eeuwijk FA (2008) A multi-trait, multi-environment QTL mixed model with an application to drought and nitrogen trials in maize (*Zea mays* L.). Euphytica 161:241–257
49. Bink MCAM, Janss LLG, Quaas RL (2000) Markov chain Monte Carlo for mapping a quantitative trait locus in outbred populations. Genet Res 75:231–241
50. Bink MCAM, Boer MP, ter Braak CJF, Jansen J, Voorrips RE, van de Weg WE (2007) Bayesian analysis of complex traits in pedigreed plant populations. Euphytica 161:85–96. doi:10.1007/s10681-007-9516-1
51. Chardon F, Virlon B, Moreau L, Falque M, Joets J, Decousset L, Murigneux A, Charcosset A (2004) Genetic architecture of flowering time in maize as inferred from quantitative trait loci meta-analysis and synteny conservation with the rice genome. Genetics 168:2169–2185
52. Jiang C, Zeng ZB (1995) Multiple trait analysis of genetic mapping for quantitative trait loci. Genetics 140:1111–1127
53. van Eeuwijk FA, Malosetti M, Boer MP (2007) Modelling the genetic basis of response curves underlying genotype x environment interaction. In: Spiertz JHJ, Struik PC, van Laar HH (eds) Scale and complexity in plant systems research. gene-plant-crop relations. Springer, Dordrecht, pp 115–126
54. Boer MP, Wright D, Feng L, Podlich DW, Luo L, Cooper M, van Eeuwijk FA (2007) A mixed-model quantitative trait loci (QTL) analysis for multiple-environment trial data using environmental covariables for QTL-by-environment interactions, with an example in maize. Genetics 177:1801–1813
55. Malosetti M, Ribaut J-M, van Eeuwijk FA (2011) The statistical analysis of multienvironment data: modelling genotype-by-environment interaction and its genetic basis. In: Drought phenotyping in crops: from theory to practice (Monneveux Philippe and Ribaut Jean-Marcel, eds). CGIAR Generation Challenge Programme, Texcoco, Mexico. In press
56. Zhang F, Zhai H-Q, Paterson AH, Xu J-L, Gao Y-M et al (2011) Dissecting genetic networks underlying complex phenotypes: the theoretical framework. PLoS ONE 6(1):e14541. doi:10.1371/journal.pone.0014541
57. Yi N, Yandell BS, Churchill GA, Allison DB, Eisen EJ, Pomp D (2005) Bayesian model selection for genome-wide epistatic quantitative trait loci analysis. Genetics 170:1333–1344
58. Xu S, Jia Z (2007) Genome wide analysis of epistatic effects for quantitative traits in barley. Genetics 176:611–623
59. Li H, Ye G, Wang J (2007) A modified algorithm for the improvement of composite interval mapping. Genetics 175:361–374
60. Li H, Ribaut J-M, Li Z, Wang J (2008) Inclusive composite interval mapping (ICIM) for digenic epistasis of quantitative traits in biparental populations. Theor Appl Genet 116: 243–260
61. Kroymann J, Mitchell-Olds T (2005) Epistasis and balanced polymorphism influencing complex trait variation. Nature 435:95–98
62. Zeng Z-B (2005) Modeling quantitative trait loci and interpretation of models. Genetics 169:1711–1725
63. Kusterer B, Muminovic J, Utz HF, Piepho H-P, Barth S, Heckenberger M, Meyer RC, Altmann T, Melchinger AE (2007) Analysis of a triple testcross design with recombinant inbred lines reveals a significant role of epistasis in heterosis for biomass-related traits in Arabidopsis. Genetics 175:2009–2017
64. Frascaroli CEMA, Landi P, Pea G, Gianfranceschi L, Villa M, Morgante M, Pè ME (2007) Classical genetic and quantitative trait loci analyses of heterosis in a maize hybrid between two elite inbred lines. Genetics 176:625–644

65. Gu X-Y, Foley ME (2007) Epistatic interactions of three loci regulate flowering time under short and long daylengths in a backcross population of rice. Theor Appl Genet 114: 745–754
66. Melchinger AE, Piepho H-P, Utz HF, Muminović J, Wegenast T, Törjék O, Altmann T, Kusterer B (2007) Genetic basis of heterosis for growth-related traits in Arabidopsis investigated by Testcross progenies of near-isogenic lines reveals a significant role of epistasis. Genetics 177: 1827–1837
67. Landers ES, Green P, Abrahamson J, Barlow A, Daly MJ, Lincoln SE, Newburg L (1987) Mapmaker: an interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. Genomics 1:174–181
68. Jansen RC (1993) Interval mapping of multiple quantitative trait loci. Genetics 135:205–211
69. Ooijen V (2004) MapQTL® 5, Software for the mapping of quantitative trait loci in experimental populations. Kyazma BV, Wageningen
70. Zeng Z-B (1994) Precision mapping of quantitative trait loci. Genetics 136:1457–1468
71. Utz HF, Melchinger AE (1996) PLABQTL: a program for composite interval mapping of QTL. J Agric Genom 2:1–5. http://probe.nalusda.gov:8000/otherdocs/jqtl/jqtl1996-01/utz.html (verified 10 September 1999)
72. Nelson JC (1997) QGene: software for marker-based genomic analysis and breeding. Mol Breed 3:229–235
73. Joehanes R, Nelson JC (2008) QGene 4.0, extensible Java QTL-analysis platform. Bioinformatics 24:2788–2789
74. Manly KF, Olson JM (1999) Overview of QTL mapping software and introduction to Map Manager QT. Mamm Genome 10:327–334
75. Portugal A, Balachandra R, Metz T, Bruskiewich R, McLaren G (2007) International crop information system for germplasm data management. In: Plant bioinformatics: methods and protocols. Humana, Totowa, pp 459–471, Chapter 22
76. McLaren CG, Metz T, van den Berg M, Bruskiewich R, Magor NP, Shires D (2009) Informatics in agricultural research for development. Adv Agron 102:135–157
77. Parkhill J, Birney E, Kersey P (2010) Genomic information infrastructure after the deluge. Genome Biol 11:402
78. Gene Ontology Consortium (2008) The Gene Ontology project in 2008. Nucleic Acids Res 36(Database issue): D440–D444
79. Avraham S, Tung CW, Ilic K, Jaiswal P, Kellogg EA, McCouch S, Pujar A, Reiser L, Rhee SY, Sachs MM, Schaeffer M, Stein L, et al (2008) The plant ontology database: A community resource for plant structure and developmental stages controlled vocabulary and annotations. Nucleic Acids Res 36(Database issue): D449–D454
80. Ilic K, Kellogg EA, Jaiswal P, Zapata F, Stevens PF, Vincent LP, Avraham S, Reiser L, Pujar A, Sachs MM, Whitman NT, McCouch SR et al (2007) The plant structure ontology, a unified vocabulary of anatomy and morphology of a flowering plant. Plant Physiol 143(2):587–599
81. Plant Ontology Consortium (2002) The Plant Ontology Consortium and plant ontologies. Comp Funct Genomics 3: 137–142
82. Bruskiewich R, Davenport G, Hazenkamp T, Metz T, Ruiz M, Simon R, Takeya M, Lee J, Senger M, McLaren G, van Hintum T (2006) The Generation Challenge Programme (GCP)—Standards for crop data. OMICS 10:215–219
83. Lee JM, Davenport GF, Marshall D, Ellis TH, Ambrose MJ, Dicks J, van Hintum TJ, Flavell AJ (2005) GERMINATE. A generic database for integrating genotypic and phenotypic information for plant genetic resource collections. Plant Physiol 139(2):619–631
84. BioMoby Consortium (2008) Interoperability with Moby 1.0—It's better than sharing your toothbrush! Brief Bioinform 9(3):220–231. doi:10.1093/bib/bbn003
85. Wilkinson M, Schoof H, Ernst R, Haase D (2005) BioMOBY successfully integrates distributed heterogeneous bioinformatics web services. The PlaNet exemplar case. Plant Physiol 138:1–13
86. Ribaut J-M, Monneveux P, Glaszmann JC, Leung H, Van Hintum T, de Vicente C (2008) International programs and the use of modern biotechnologies for crop improvement. In: Moore P, Ming R (eds) Genomics of tropical crop plants. Springer, New York, pp 21–63
87. Sonnino A, Carena MJ, Guimarães EP, Baumung R, Pilling D, Rischkowsky B (2007) An assessment of the use of molecular markers in developing countries. In: Guimarães EP, Ruane J, Scherf BD, Sonnino A, Dargie JD (eds) Marker-assisted selection: Current status and future perspectives in crops, livestock, forestry and fish. FAO, Rome, pp 15–26
88. Huang J, Rozelle S, Pray C, Wang Q (2002) Plant biotechnology in China. Science 295:674–677
89. Suresh P, Devi SV, Choudhary UN (2008) Resources and priorities for plant biotechnology research in India. Curr Sci 95:1400–1402
90. Ghneim Herrera T, Posso Duque D, Pérez Almeida I, Torrealba Nuñez G, Pieters AJ, Martínez CP, Tohme JM (2008) Assessment of genetic diversity in Venezuelan rice cultivars using simple sequence repeats markers. Electron J Biotechnol. doi:10.2225/vol11-issue5-fulltext-6
91. Khadari B, Oukabli A, Ater M, Mamouni A, Roger JP, Kjellberg F (2004) Molecular characterization of Moroccan fig germplasm using intersimple sequence repeat and simple sequence repeat markers to establish a reference collection. Hortic Sci 40:29–32
92. Onguso JM, Kahangi EM, Ndiritu DW, Mizutani F (2004) Genetic characterization of cultivated bananas and plantains in Kenya by RAPD markers. Sci Hortic 99:9–20
93. Paredes M, Becerra V, González MI (2008) Low genetic diversity among garlic (*Allium sativum* L.) accessions detected using random amplified polymorphic DNA (RAPD). Chil J Agric Res 68:3–12
94. Abalo G, Tongoonaa P, Derera J, Edema R (2009) A comparative analysis of conventional and marker-assisted selection methods in breeding maize streak virus resistance in maize. Crop Sci 49:509–520

95. Danson JW, Mbogori M, Kimani M, Lagat M, Kuria A, Diallo A (2006) Marker-assisted introgression of opaque2 gene into herbicide-resistant elite maize inbred lines. Afr J Biotechnol 5:2417–2422
96. Okogbenin E, Porto MCM, Egesi C, Mba C, Espinosa E, Santos LG, Ospina C, Marin J, Barrera E, Gutierrez J et al (2007) Marker-assisted introgression of resistance to cassava mosaic disease into Latin American germplasm for the genetic improvement of cassava in Africa. Crop Sci 47:1895–1904
97. Leung H, Wu J, Liu B, Bustaman M, Sridhar R, Singh K, Redona E, Quang VD, Zheng K, Bernardo M et al (2004) Sustainable disease resistance in rice: current and future strategies. In: New directions for a diverse planet. Proceedings of the 4th international crop science congress, 26 September–1 October, Brisbane
98. Sagredo B, Mathias M, Barrientos C, Acuña I, Kalazich J, Santosrojas J (2009) Evaluation of a SCAR RYSC3 marker of the RYadg gene to select resistant genotypes to potato virus Y (PVY) in the INIA potato breeding program. Chil J Agric Res 69:305–315
99. Stevens R (2008) Prospects for using marker-assisted breeding to improve maize production in Africa. J Sci Food Agric. doi:10.1002/jsfa.3154
100. Hartwich F, Tola J, Engler A, González C, Ghezan G, Vázquez-Alvarado JMP, Silva JA, Espinoza JJ, Gottret MV (2007) Building public–private partnerships for agricultural innovation, Food security in practice technical guide series. International Food Policy Research Institute, Washington, DC
101. Jaccoud D, Peng K, Feinstein D, Kilian A (2001) Diversity arrays: a solid state technology for sequence information independent genotyping. Nucleic Acids Res 29:e25
102. Ganal MW, Altmann T, Roder M (2009) SNP identification in crop plants. Curr Opin Plant Biol 12:211–217
103. Varshney RK, Penmetsa RV, Dutta S, Kulwal PL, Saxena RK, Datta S, Sharma TR, Rosen B, Carrasquilla-Garcia N, Farmer A et al (2009) Pigeonpea genomics initiative (PGI): an international effort to improve crop productivity of pigeonpea (*Cajanus cajan* L.). Mol Breed 26:393–408. doi:10.1007/s11032-009-9327-2
104. Varshney RK, Close TJ, Singh NK, Hoisington DA, Cook DR (2009) Orphan legume crops enter the genomics era! Curr Opin Plant Biol 12:1–9
105. Ajani EN, Madukwe MC, Agwu AE, Onwubuya EA (2009) Assessment of technology generating institutions in biotechnology innovation system of South-Eastern Nigeria. Afr J Biotechnol 8:2258–2264
106. O'Toole JC, Toenniessen GH, Murashige T, Harris RR, Herdt RW (2001) The Rockefeller Foundation's international program on rice biotechnology. In: Khush GS, Brar DS, Hardy B (eds) Rice genetics IV. Proceedings of the 4th international rice genetics symposium, Los Baños. International Rice Research Institute, pp 39–59
107. Kelemu S, Mahuku G, Fregene M, Pachico D, Johnson N, Calvert L, Rao I, Buruchara R, Amede T, Kimani P et al (2003) Harmonizing the agricultural biotechnology debate for the benefit of African farmers. Afr J Biotechnol 2:394–416
108. Morris M, Edmeades G, Peju E (2006) The global need for plant breeding capacity: what roles for the public and private sectors? Hortic Sci 41:30–39
109. Salekdeh GH, Reynolds M, Bennett J, Boyer J (2009) Conceptual framework for drought phenotyping during molecular breeding. Trends Plant Sci 14:488–496
110. Hyman G, Fujisaka S, Jones P, Wood S, de Vicente C, Dixon J (2008) Strategic approaches to targeting technology generation: assessing the coincidence of poverty and drought-prone crop production. Agric Syst 98:50–61
111. Hamer G, Cooper M, Tardieu F, Welch S, Walsh B, van Euuwijk F, Chapman S, Polish D (2006) Models for navigating biological complexity in breeding improved crop plants. Trends Plant Sci 11:587–593
112. Ribaut J-M, Betran J, Monneveux P, Setter T (2008) Drought tolerance in maize. In: Bennetzen J, Hake S (eds) Maize handbook, vol 1. Springer, New York, pp 311–344
113. Cooper M, van Eeuwijk F, Hammer GL, Podlich DW, Messina C (2009) Modeling QTL for complex traits: detection and context for plant breeding. Curr Opin Plant Biol 12:231–240
114. Mackay TFC, Stone EA, Ayroles JF (2009) The genetics of quantitative traits: challenges and prospects. Nat Rev Genet. doi:10.1038/nrg2612
115. Cooper M, van Eeuwijk FA, Chapman SC, Podlich DW, Löffler C (2006) Genotype-by-environment interactions under water-limited conditions. In: Ribaut JM (ed) Drought adaptation in cereals. Haworth, Binghampton, pp 51–95
116. Chagné D, Batley J, Edwards D, Forster JW (2007) Single nucleotide polymorphism genotyping in plants. In: Oraguzie NC, Rikkerink EHA, Gardiner SE, de Silva HN (eds) Association mapping in plants. Springer, New York, pp 77–94
117. Angaji SA (2009) Single nucleotide polymorphism genotyping and its application on mapping and marker-assisted plant breeding. Afr J Biotechnol 8:908–914
118. Muchero M, Diop NN, Bhat PR, Fenton RD, Wanamaker S, Pottorff M, Hearne S, Cisse N, Fatokun C, Ehlers JD et al (2009) A consensus genetic map of cowpea [*Vigna unguiculata* (L) Walp.] and synteny based on EST-derived SNPs. Proc Natl Acad Sci USA 106:18159–18164
119. Kawuki RS, Ferguson M, Labuschagne M, Herselman L, Kim DJ (2009) Identification, characterisation and application of single nucleotide polymorphisms for diversity assessment in cassava (*Manihot esculenta* Crantz). Mol Breed 23:669–684
120. Dwivedi SL, Crouch JH, Mackill DJ, Xu Y, Blair MW, Ragot M, Upadhyaya HD, Ortiz R (2007) The molecularization of public sector crop breeding: progress, problems, and prospects. Adv Agron 95:163–318. doi:10.1016/S0065-2113(07)95003-8
121. Xu Y, Crouch JH (2008) Marker-assisted selection in plant breeding: from publications to practice. Crop Sci 48:391–407
122. Dreher K, Khairallah M, Ribaut J-M, Morris M (2003) Money matters (I): costs of field and laboratory procedures associated with conventional and marker-assisted maize breeding at CIMMYT. Mol Breed 11:221–234

123. Morris M, Dreher K, Ribaut J-M, Khairallah M (2003) Money matters (II): costs of maize inbred line conversion schemes at CIMMYT using conventional and marker-assisted selection. Mol Breed 11:235–247
124. Delmer DP (2005) Agriculture in the developing world: connecting innovations in plant research to downstream applications. Proc Natl Acad Sci USA 102:15739–15746
125. Guimarães EP, Kueneman E, Carena MJ (2006) Assessment of national plant breeding and biotechnology capacity in Africa and recommendations for future capacity building. Hortic Sci 41:50–52
126. Ismail AM, Heuer S, Thomson MJ, Wissuwa M (2007) Genetic and genomic approaches to develop rice germplasm for problem soils. Plant Mol Biol 4:547–570
127. Alpuerto VE, Norton GW, Alwang J, Ismail AM (2009) Economic impact analysis of marker-assisted breeding for tolerance to salinity and phosphorous deficiency in rice. Rev Agr Econ 31:779–792
128. Septiningsih EM, Pamplona AM, Sanchez DL, Neeraja CN, Vergara GV, Heuer S, Ismail AM, Mackill DJ (2009) Development of submergence-tolerant rice cultivars: the Sub1 locus and beyond. Ann Bot 103:151–160

Molten Carbonate Fuel Cells

Choong-Gon Lee
Department of Chemical Engineering, Hanbat
National University, Yuseong-gu, Daejeon,
South Korea

Article Outline

Glossary

Activation overpotential Voltage loss due to low charge-transfer rate on the electrode surface.

Anode A porous electrode where hydrogen is oxidized with carbonate ions ($CO_3{}^{2-}$) to steam and carbon dioxide.

Basicity Log(k_d) of molten carbonates where k_d is the equilibrium constant of the reaction $CO_3^{2-} \overset{k_d}{\rightleftharpoons} O^{2-} + CO_2$ similar to the pH of aqueous solutions.

Cathode A porous electrode where oxygen is reduced with carbon dioxide to carbonate ions ($CO_3{}^{2-}$).

Electrolyte Molten carbonates providing ionic paths for the electrode reactions with combinations of Li_2CO_3, Na_2CO_3, and K_2CO_3.

Exchange current density ($\mathbf{i_o}$) An actual current density of an electrode at net zero current indicating catalytic activity of the electrode.

Fuel cell A system of continuous electrochemical energy conversion from chemical energy to electricity mostly by oxidation of hydrogen and reduction of oxygen.

Internal resistance Electrical resistance of cell components.

Mass transfer Access of reactants to the electrode surface and departure of products mainly by diffusion and convection.

Matrix Ceramic porous material holding molten carbonates by capillary forces.

Ohmic loss ($\boldsymbol{\eta}_{\mathbf{IR}}$) Voltage loss due to electrical resistance of cell components.

Open circuit voltage ($\mathbf{E_{OCV}}$) A cell voltage at net zero current determined by the relation of $E_{OCV} = E^o + \frac{RT}{2F}\ln\left(\frac{p(H_2)p(O_2)^{0.5}p(CO_2)_{ca}}{p(H_2O)p(CO_2)_{an}}\right)$.

Overpotential ($\boldsymbol{\eta}$) Voltage reduction from an open circuit voltage due to resistance of electrochemical reactions at an electrode.

Polarization A state of deviation from open circuit voltage due to current flowing in the cell.

Reaction kinetics Charge-transfer rates of electrochemical reactions on electrode surfaces.

Three-phase boundary A site of electrode-carbonate electrolyte-gaseous reactants where electrochemical reactions take place.

Definition of the Subject

Two parts are treated: one is the physical and chemical features of materials of molten carbonate fuel cells (MCFCs), and the other is performance analysis with a 100 cm^2 class single cell. The characteristics of the fuel cell are determined by the electrolyte. The chemical and physical properties of the electrolyte with respect to gas solubility, ionic conductivity, dissolution of cathode material, corrosion, and electrolyte loss in the real cell are introduced. The reaction characteristics of hydrogen oxidation in molten carbonates and materials for the anode of the MCFC are reviewed. The kinetics of the oxygen reduction reaction in the molten carbonates and state of the art of cathode materials are also described. Based on the reaction kinetics of electrodes, a performance analysis of MCFCs is introduced. The performance analysis has importance with respect to the increase in performance through material development and the extension of cell life by cell development. Conventional as well as relatively new analysis methods are introduced.

Introduction

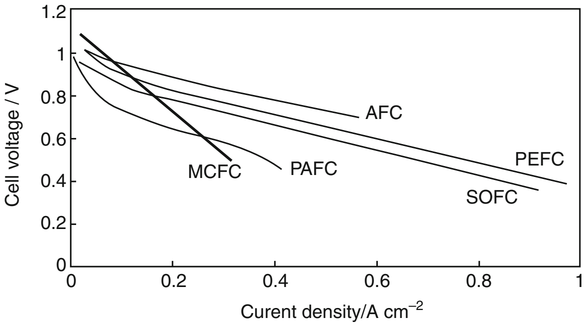
A fuel cell, as an emerging power source, generates power directly from the chemical energy of fuel. The fuel cell runs with electrochemical reactions at the electrodes where mostly H_2 oxidation at the anode, and O_2 reduction at the cathode occur. The power generation scheme of a fuel cell is very different from that of conventional grid power, where the electricity comes from electromagnetic induction with mechanical rotation. Thus, conventional power is produced by several steps of chemical, mechanical, and electrical energy changes. In contrast, a fuel cell converts chemical energy to electrical energy directly, which allows high energy conversion efficiency and low pollution emission from the fuel cell.

Among the fuel cells, the PEMFC (Polymer Electrolyte Membrane Fuel Cell or PEFC), AFC (Alkaline Fuel Cell), and PAFC (Phosphoric Acid Fuel Cell) are run by H^+ and OH^- movements in the electrolyte. On the other hand, the MCFC (Molten Carbonate Fuel Cell) and SOFC (Solid Oxide Fuel Cell) work with carbonate ions ($CO_3{}^{2-}$) and oxide ions (O^{2-}), respectively. The acid (H^+) and base (OH^-) generally run up to 200°C whereas carbonate and oxide ions run at over 600°C. Thus, the PEMFC, AFC, and PAFC have relatively low operating temperatures compared with the MCFC and SOFC.

The MCFC, running with molten carbonate electrolytes, has an operating temperature of about 650°C. This high temperature facilitates electrochemical reactions at the electrodes, allowing inexpensive metal electrodes such as Ni to be used while low temperature fuel cells, that is, PEMFC, AFC, and PAFC, require Pt electrocatalysts. The high operating temperature also gives high efficiency through the bottoming cycle that utilizes exhaust heat. Thus, the MCFC has some merits of high efficiency and system economics over the low temperature fuel cells.

Figure 1 shows current-voltage behaviors of fuel cells. The polarization behavior represents the reaction and performance characteristics of the fuel cells. In principle, the low temperature fuel cell has a larger absolute Gibbs free energy of H_2O formation. Thus, the PEMFC, AFC, and PAFC have a higher open circuit voltage (E_{OCV}) than the MCFC and SOFC. However, the low temperature fuel cells also have a sluggish charge-transfer rate in the electrode reaction, which leads to high activation overpotential. In general, the activation overpotential exponentially decreases voltage due to the applied currents at the early stage of current application (Fig. 1). Moreover, the Pt electrode in a low temperature fuel cell is oxidized at open circuit voltage. Therefore, low temperature fuel cells show an exponential decrease in voltage at around zero current and unclear E_{OCV} values. On the other hand, the MCFC shows a monotonic decrease in voltage even at the early stage of current application. This indicates that the electrode reactions have very small activation overpotential, probably due to the high temperature molten carbonate electrolytes. However, the steep current-voltage behavior of the MCFC shows that it has relatively high internal resistance and electrochemical reaction resistance in the cell compared with other fuel cells.

The carbonate ions ($CO_3{}^{2-}$) are supplied by the electrolytes: a combination of Li_2CO_3, K_2CO_3, and Na_2CO_3. Current electrolytes for the MCFC are the



Molten Carbonate Fuel Cells. Figure 1
Current-voltage behaviors of various hydrogen fuel cells (From [46])

eutectics of Li_2CO_3-K_2CO_3 and Li_2CO_3-Na_2CO_3, which have melting temperatures of about 500°C. Then the electrode reactions with H_2 and O_2 in the MCFC are as follows:

$$\text{Anode}\ H_2 + CO_3^{2-} \rightarrow H_2O + CO_2 + 2e^- \tag{1a}$$

$$\text{Cathode}\ \frac{1}{2}O_2 + CO_2 + 2e^- \rightarrow CO_3^{2-} \tag{1b}$$

$$\text{Total}\ H_2 + \frac{1}{2}O_2 \rightarrow H_2O + \text{Heat} + \text{Power} \tag{1c}$$

The molten carbonates also allow the use of CO as a fuel. Then the reactions are as follows:

$$\text{Anode}\ CO + CO_3^{2-} \rightarrow 2CO_2 + 2e^- \tag{2a}$$

$$\text{Cathode}\ \frac{1}{2}O_2 + CO_2 + 2e^- \rightarrow CO_3^{2-} \tag{2b}$$

$$\text{Total}\ CO + \frac{1}{2}O_2 \rightarrow CO_2 \tag{2c}$$

This is a very unique characteristic of a high temperature fuel cell because CO behaves as a fuel in the MCFC while it is a poisonous species for the low temperature fuel cells. It gives fuel diversity to the MCFC. At present, the H_2 fuel is supplied by methane steam reforming as follows:

$$CH_4 + H_2O \rightarrow CO + 3H_2 \tag{3a}$$

$$CO + H_2O \rightarrow CO_2 + H_2 \tag{3b}$$

Methane is a main component of natural gas (over 90 vol.%) and its infrastructure is relatively well built across the world. Thus, natural gas is a main source of H_2 fuel for fuel cells. However, natural gas is much more expensive than coal, and thus fuel cells are economically inferior to coal power. Coal based operations have potential to improve the economics of fuel cells. At present, coal gasification is utilized as a clean coal technology, producing H_2 and CO gases as its main components. Organic materials are also generally decomposed to H_2 and CO. Thus, MCFCs can run with coal and organic wastes, which will enhance the economics of MCFCs.

In the past 35 years, MCFCs have been developed in the world. The primary developer is FCE (Fuel Cell Energy Co.) in the USA, which has developed an internal reforming type MCFC unit stack of up to 300 kW, and provides up to 2.4 MW MCFC systems by combination of the unit stacks. It also reported that a combination of an MCFC and a gas turbine recorded 56% electrical efficiency (based on the lower heating value (LHV) of natural gas) [1]. MTU, a German company, has designed a 250 kW class system, "Hot Module," by adapting FCE's stack. Japan developed MCFC systems of up to 1 MW class during the past 30 years. The 1 MW system was the pressurized external reforming type, and IHI (Ishikawajima-Harima Heavy Industry) produced a 300 kW class external reforming type system. Ansaldo Co. of Italy has developed a 500 kW class MCFC system. A 300 kW class MCFC system is also under development in Korea.

In this work, the characteristics of material for the MCFC such as the electrolyte, electrodes, and matrix are introduced. In addition, the diagnostic tools for MCFC performance are also treated: conventional methods of steady-state polarization, current interruption (C/I), and AC impedance, and novel methods of inert gas step addition (ISA) and reactant gas addition (RA).

Components of MCFC

Electrolytes

The MCFC is an electrochemical reaction system where the anode oxidizes H_2 to H_2O and the cathode reduces O_2 to ${CO_3}^{2-}$ as shown in Eqs. 1a and 1b. Thus carbonate materials serve as the electrolyte, which is generally a mixture of various alkali metal carbonates of Li_2CO_3, Na_2CO_3, and K_2CO_3. Table 1 shows the melting points (m.p.), surface tension (γ), density (ρ), electric conductivity (κ), and Henry's Law constant of O_2 dissolution (h_{O2}) for various eutectic carbonates.

The table also shows the properties of an aqueous solution (1 M, KCl). It is found that the eutectics are a kind of electrolyte that have about twice the density, about three times the surface tension, and over ten times the electric conductivity of an aqueous solution. Eutectics also have much lower melting points than at single carbonates: Li_2CO_3 (999 K), Na_2CO_3 (1,131 K), and K_2CO_3 (1,172 K).

In general, the molten carbonates have an equilibrium with oxide ions (O^{2-}) and CO_2.

$$CO_3^{2-} \overset{k_d}{\rightleftharpoons} O^{2-} + CO_2 \tag{4}$$

where k_d is the dissociation constant $(= [O^{2-}] \cdot p(CO_2))$. Molten carbonates are also

Molten Carbonate Fuel Cells. Table 1 Properties of eutectic alkali metal carbonates [2]

System	Composition (mol%)	m.p. (K)	γ (mN/m) at 973 K	ρ (g/cm^3) at 973 K	κ (S/cm) at 973 K	h_{O2} (mol/cm^3 atm) at 923 K
Li_2CO_3:Na_2CO_3	53:47		239.0	1.937	2.181	1.83×10^{-7}
	52:48	774				
Li_2CO_3:K_2CO_3	62:38	761	214.1	1.912	1.053	
	50:50	777.5	206.0	1.917		3.26×10^{-7}
Li_2CO_3:Na_2CO_3:K_2CO_3	43.5:31.5:25	670	219.6	1.984	1.476	3.91×10^{-7}
Na_2CO_3:K_2CO_3	56:44	983				
1 M KCl at 298 K [48]			~72[a]		0.108	

[a] Value for water

Molten Carbonate Fuel Cells. Table 2 Dissociation constants (k_d) of carbonate melts [4]

Composition	k_d		
	823 K	923 K	1,023 K
Li_2CO_3	1.01×10^{-6}	2.08×10^{-5}	2.37×10^{-4}
53 mol% Li_2CO_3 – 47 mol% Na_2CO_3	2.58×10^{-9}	1.14×10^{-7}	2.41×10^{-6}
43.5 mol% Li_2CO_3 – 31.5 mol% Na_2CO_3 – 25.0 mol% K_2CO_3	2.04×10^{-10}	1.23×10^{-8}	3.34×10^{-7}
50 mol% Li_2CO_3 – 50 mol% K_2CO_3	9.77×10^{-11}	6.43×10^{-9}	1.86×10^{-7}
56 mol% Na_2CO_3 – 44 mol% K_2CO_3	8.53×10^{-14}	1.38×10^{-11}	8.26×10^{-10}

existing in the form of alkali metal cations (M^{2+}) and carbonate anions ($CO_3{}^{2-}$). Thus, the activity of oxide ions in the melt determines the characteristics of the melt as the activity of protons (H^+) in the aqueous solution (pH) represents the acidity of the solution. By adopting the Lux–Flood acid-base theory, the basicity of the molten carbonate can be defined.

In aqueous solution: acid = base + H^+
In carbonate melts: base = acid + O^{2-}

Therefore, the activity of oxide ions indicates the basicity of the melt and CO_2 in the melt behaves as an acid.

Table 2 shows the dissociation constants (k_d) of alkali carbonate melts. The constant rises with Li content and decreases with K and Na contents. Thus the melts become more basic with increasing Li content. The behavior is in agreement with the ionic radius of alkali metals: the smaller ionic radius of alkali metal results in more basic melts. Then the series of basic melts is as follows:

$$\begin{aligned} Li_2CO_3 > Li_2CO_3 - Na_2CO_3 > Li_2CO_3 \\ - K_2CO_3 > Na_2CO_3 - K_2CO_3 \end{aligned}$$

The oxygen solubility also becomes higher in more acidic carbonate melts. The behavior can be interpreted from the ionic radius. The larger ionic radius of alkali metal enhances O_2 solubility. This is due to the increased stability of superoxide ions ($O_2{}^-$) and peroxide ions ($O_2{}^{2-}$) in the carbonate melts, which are produced by the following chemical reactions:

$$O_2 + 2CO_3^{2-} \rightleftharpoons 2O_2^{2-} + 2CO_2 \tag{5a}$$

$$3O_2 + 2CO_3^{2-} \rightleftharpoons 4O_2^{-} + 2CO_2 \tag{5b}$$

The dissociation constants also rise with temperature, indicating that the melts become more basic at a higher temperature.

As another property of molten carbonates, a contact angle is to be introduced. The molten carbonates are transparent liquid and movable. To maintain the molten carbonates in the fuel cell, a porous matrix structure is required. The porous structure holds the carbonate melts by the capillary forces of the pores. The carbonate melts have a very low contact angle with oxide (≈ 0°) but very large contact angles with metals [2]. Thus, the cathode of the oxide electrode is very well wetted by the carbonates, while the anode electrode, with a metal state under strong reductants of H_2, is poorly wetted. Based on this concept, a dry agglomerate model for the anode and a well wetted agglomerate model for the cathode have been suggested [3]. According to the different surface tensions of Li-K and Li-Na carbonate melts as shown in Table 1, the melts have different wetting behaviors. The high surface tension of Li-Na carbonate melts causes poor wetting behavior compared with Li-K melts. This results in a steeper overpotential increase in the Li-Na carbonate electrolyte cell than in the Li-K cell at 600°C [4, 5].

Gas solubility is also an important parameter for the electrode reaction because the electrode is covered by carbonates and gas reactants must transfer through the carbonate film. So, higher solubility results in less kinetic and mass-transfer resistances during the electrode reaction. In general, CO_2 and H_2 solubilities are approximately 10^{-5} mole cm^{-3} atm^{-1} and the O_2 solubility is about one tenth of the H_2 solubility in the carbonate melts. The following dependence of gas solubility on the melt compositions at the same gas condition and temperature was reported [6]:

O_2 solubility: Li-K $>$ Li-Na $\approx$ Li-Na-K
H_2 solubility: Li-K $>$ Li-Na-K
CO_2 solubility: Li-Na $>$ Li-Na-K $>$ Li-K

Electrolyte Loss in the Cell

The carbonate electrolytes are contained in the matrices, which are porous ceramic materials placed between the anode and cathode. Since the matrix is comprised of sub-micron size pores, most of the pores are filled with carbonate electrolytes. Thus, the electrolyte prevents gas leakage between the electrodes. The electrodes are covered by metal separators that provide gas flow paths over the electrodes; the peripheral area of the separators is sealed by matrices and is called the wet seal area. Thus, the matrix prevents gas leaks from the inside to the outside of the cell by wet sealing.

Electrolyte loss weakens the gas sealing and shortens cell life. The molten carbonates are very corrosive materials. Since the separator is made of stainless steel, corrosion takes place on the surface of the separator. To reduce the corrosion, the anode side is coated with Ni and the wet seal area with Al.

According to the report on a 2 MW field test at Santa Clara, the causes and amounts of electrolyte losses are as follows: cathode hardware loss, 73%; fixed losses, 17%; vaporization loss, 7%; and unaccounted loss, 3% [7]. The report pointed out that most of the electrolyte loss was due to corrosion at the cathode because the Ni and Al coatings on the anode and wet seal area respectively prevented serious corrosion. Mitsubishi Electric Co. reported that about half of the electrolyte loss was due to the corrosion at the cathode current collector and the loss was proportional to the area of the current collector [8].

Among the austenitic stainless steels, 316 L and 310 are generally used for the MCFC separator and current collector. The Cr contents of 316 L and 310 are about 17% and 25%, respectively, and thus a denser $LiCrO_2$ layer occurs on the surface of 310 with more corrosion resistivity. However, $LiCrO_2$ has very low conductivity, resulting in a high electrical resistance with 310. Moreover, 310 makes water soluble K_2CrO_4 as a corrosion product, and thus electrolyte loss is more severe with 310 than 316 L [9].

Li-K and Li-Na carbonate melts had different corrosion behaviors with 316 L. These melts showed insignificant corrosion at 650°C but severe pit corrosion was observed only in the Li-Na melt under the present O_2 and CO_2 condition at around 550°C [10, 11]. Mitsubishi Electric Co. reported that an inert gas condition in the temperature range could prevent severe corrosion [10]. IHI in Japan also reported that the corrosion is mitigated by pre-oxidation of the surface by steam [11].

The molten carbonates also react with steam, which produces hydroxides, LiOH, NaOH, and KOH. It was reported that KOH has about twice the vapor pressure of LiOH and the highest vapor pressure among the above hydroxides [12]. However, as reported for the Santa Clara test [7], the electrolyte loss by vaporization was 7% of total loss, which is still low compared with the total loss.

Anode

In general, H_2 oxidation is a quite fast reaction even in the low temperature fuel cells. The H_2 oxidation in the high temperature molten carbonate is fast enough, as expected. Among the several mechanisms, the following two are mainly discussed.

Ang and Sammells' mechanism [13]

$$\begin{gathered} H_2 + 2M \rightleftharpoons 2M - H \\ M - H + CO_3^{2-} \rightarrow OH^- + CO_2 + M + e^- \\ M - H + OH^- \rightleftharpoons H_2O + M + e^- \end{gathered} \tag{6}$$

Suski's mechanism [14]

$$\begin{gathered} H_2 + 2M \rightleftharpoons 2M - H \\ 2\{M - H + CO_3^{2-} \rightarrow OH^- + CO_2 + M + e^-\} \\ 2OH^- + CO_2 \rightleftharpoons H_2O + CO_3^{2-} \end{gathered} \tag{7}$$

The mechanism of Ang and Sammells was suggested on the basis of experimental results obtained with Ni [13] and Cu [15] electrodes. The other mechanism was supported by the role of OH^- [14], one electron number and double the rate determining step [16]. Although the reaction mechanisms are not yet in agreement, the kinetic values are within an acceptable range: Ni has an exchange current density of about 100 mA cm^{-2} at 923 K in Li-K melts. Moreover, the following reaction orders are in agreement with the experimental results:

$$i_o = i_o^o p(H_2)^{0.25} p(CO_2)^{0.25} p(H_2O)^{0.25} \tag{8}$$

The positive reaction orders of CO_2 and H_2O reflects the fact that CO_2 and H_2O in the anode enhance the reaction kinetics although they are product species.

Owing to the high temperature, H_2 oxidation is largely insensitive to the chosen anode materials. The exchange current densities of Co [13], Cu [15], Pt [16], Ir [16], Au [16], and Ag [16] were about 45, 69, 85, 27, 26, and 19 mA cm^{-2}, respectively at 923 K in Li-K carbonate melts. Considering the possible experimental error, these values indicate that the materials have a similar H_2 oxidation rate at the anode. In addition, the oxidation potential of Ni in molten carbonate is about −0.802 V under a 1 atm O_2 condition at 925 K [2]. Thus, Ni is metallic at open circuit voltage, and is not oxidized up to a certain potential.

The electrodes in the fuel cell should provide solid-liquid-gas three-phase boundary to reduce overpotential. Porous type electrodes are designed and the carbonate electrolytes are dispersed in the electrode by capillary forces. The anode has a higher contact angle and lower wetting with carbonates than the cathode, which allows a smaller pore size at the anode. As mentioned above, the anode has a very high H_2 oxidation rate. So, the active surface area and electrolyte filling in the anode are not critical parameters for its performance. Therefore the anode behaves as an electrolyte reservoir in the MCFC.

The most plausible life-limiting factor is electrolyte depletion in the MCFC. Since the anode serves as electrolyte reservoir, its stable pore structure is very important for the electrolyte management. The Ni-Cr alloy electrodes, however, showed creep behavior, which deforms the pore structure due to Ostwald ripening under the suppression in the carbonate melts. Thus, the anode thickness is reduced with time and electrolyte amounts in the anode are also decreased. Consequently, the cell life can be shortened. It was reported that the addition of Al to Ni-Cr resulted in higher creep resistivity than in Ni-Cr [12].

As alternative anodes, Cu-Al alloy and $LiFeO_2$ were tested, but they had insufficient creep strengths [1].

Cathode Electrode

Oxygen reduction is generally much slower than H_2 oxidation. In particular, low temperature fuel cells have much larger overpotential at the cathode, which is limiting the cell performance. A similar tendency also prevails in MCFCs and most researches on kinetics have focused on oxygen reduction. In the past 35 years, a lot of works on oxygen reduction in molten carbonates have been done and several oxygen reduction mechanisms have been suggested. Among them, two mechanisms, superoxide and peroxide paths, have been mainly suggested from half-cell experiments with plain gold electrodes [17, 18].

Superoxide path [17]

$$3O_2 + 2CO_3^{2-} \overset{k}{\rightleftharpoons} 4O_2^- + 2CO_2 \tag{9a}$$

$$O_2^- + e^- \rightarrow O_2^{2-} \tag{9b}$$

$$O_2^{2-} + 2CO_2 + 2e^- \rightarrow 2CO_3^{2-} \tag{9c}$$

$$i_o = i_o^o p(O_2)^{0.625} p(CO_2)^{-0.75} \tag{9d}$$

Peroxide path [18]

$$O_2 + 2CO_3^{2-} \rightleftharpoons 2O_2^{2-} + 2CO_2 \tag{10a}$$

$$O_2^{2-} + 2CO_2 + 2e^- \rightarrow 2CO_3^{2-} \tag{10b}$$

$$i_o = i_o^o p(O_2)^{0.375} p(CO_2)^{-1.25} \tag{10c}$$

where i_o is the exchange current density, which represents the reaction rate on the electrode surface, and i_o^o is the intrinsic exchange current density. The superoxide path was observed mostly in the acidic carbonate melts of Na-K, whereas the peroxide path was found in the most basic melt of Li_2CO_3. In these melts, general agreements were obtained. However, the two mechanisms were separately suggested by the researchers for the widely used Li-K carbonate melts. This ambiguity probably resulted from experimental difficulties with the very hot corrosive carbonate melts. Among the kinetic works on oxygen reduction, the very high exchange current density in oxygen reduction ($i_o \approx 10$ mA cm^{-2}) was generally accepted. The value is about eight orders higher than that of O_2 reduction in an aqueous solution ($i_o \approx 10^{-7}$ mA cm^{-2} [19]). This very high value allows the expectation of very small activation overpotential at the cathode in the MCFC.

The oxygen reduction characteristics were also investigated with 100 cm^2 class single cells by Japanese and Korean groups. In particular, Uchida's group has suggested that oxygen reduction in Li-K melts is a process of mixed diffusion of superoxide and CO_2 in the melts [20].

$$\begin{aligned} R_{ca,L} &= R_{O_2^-} + R_{CO_2} \\ &= \frac{RTd}{3^2 F^2 D_{O_2^-} \kappa p(O_2)^{0.75} p(CO_2)^{-0.5}} \\ &\quad + \frac{RTdh_{CO_2}}{1.5^2 F^2 D_{CO_2} p(CO_2)} \end{aligned} \tag{11a}$$

$$R_{ca,L} p(CO_2) = Ap(O_2)^{-0.75} p(CO_2)^{1.5} + B \tag{11b}$$

where $R_{ca,L}$ is the sum of the diffusion resistance of superoxide ions and CO_2, A and B are constants, $A = \frac{RTd}{3^2 F^2 D_{O_2^-} \kappa}$, $B = \frac{RTdh_{CO_2}}{1.5^2 F^2 D_{CO_2}}$, D is the diffusivity, F is the Faraday constant, κ is the equilibrium constant of Reaction Eq. 9a, d is the film thickness, and h is the Henry constant. The diffusion resistance ($R_{ca,L}$) is the main part of the overpotential, and thus kinetic analysis is available with a full cell. They analyzed cathodic overpotential with respect to the O_2 and CO_2 gas partial pressures and consistently concluded that the superoxide mechanism prevails under a normal condition of the Li-K carbonate single cells [21, 22].

Under the oxidizing conditions of the cathode, oxides are mostly stable. It was suggested that Au, NiO, and SnO_2 have exchange current densities (i_o) of 38.5, 18.3, and 11.2 mA cm^{-2}, respectively, indicating that material species are insignificant for the O_2 reduction rate [23]. Since NiO shows comparable catalytic behavior to Au, NiO is the most popular cathode material so far. In general, Ni is oxidized to NiO inside the cell during the pretreatment procedure of MCFC, which is in situ oxidation. During the oxidation, Li ions in the carbonate melts are doped into the NiO. Reportedly about 2% of Li is doped, and the Li doped NiO has about 33 S cm^{-1} electronic conductivity, which is close to metal conductivity [24].

However, NiO at the cathode dissolves into the carbonate electrolyte (Eq. 12), and the Ni ions are reduced to Ni metal in the matrix by H_2 from the anode. Consequently, Ni deposition in the matrix may cause electrical short circuit between the anode and cathode.

$$NiO + CO_2 = Ni^{2+} + CO_3^{2-} \tag{12}$$

It was reported that NiO dissolution is proportional to CO_2 partial pressure and activity [25]. Thus acidic melts have higher NiO solubility. Several methods have been employed to reduce the NiO dissolution: (1) increased basicity of the melts, (2) finding an alternative cathode material, and (3) modification of the NiO electrode. For the first approach, Li-Na melts were considered instead of widely used Li-K melts. It was reported that Li-Na melts have lower solubility [25] and higher electric conductivity than Li-K melts, as shown in Table 1. A Japanese group also reported that a Li-Na carbonate electrolyte cell showed higher performance than a cell with Li-K carbonates in the temperature range 575–675°C up to 5 atm pressure [5]. The addition of alkali earth metals, MgO, BaO, SrO, and CaO, to the carbonate melts was attempted because alkali earth metal ions behave as a strong base in the melts [26]. They definitely reduced the NiO solubility. However, alkali earth metal ions were segregated in the matrix and the metal ions were distributed at the anode side. In conclusion, the addition was not effective [27]. As a second approach, oxides such as $LiCoO_2$ and $LiFeO_2$ were developed for the cathode material. Although the solubility of $LiCoO_2$ is about one third of that of NiO [28], the low conductivity and high cost of $LiCoO_2$ are obstructions to its use. $LiFeO_2$ also has low solubility compared with NiO but very low conductivity is also a barrier [29]. As a third method, modification of NiO with MgO and Fe_2O_3 was also attempted. The NiO-MgO-$LiFeO_2$ is a solid solution and the material has lower NiO solubility due to the stable structure [30]. Industrially, thickening the matrix and reducing the CO_2 partial pressure were attempted, and FCE Co. reported that it could ensure a cell life of 5 years using these methods [1].

Performance Analysis

The thermodynamic electromotive force of MCFC, called the open circuit voltage (E_{OCV}), is determined by the following equation according to Eqs. 1a and 1b:

$$E_{OCV} = E^{o} + \frac{RT}{2F}\ln\left(\frac{p(H_2)p(O_2)^{0.5}p(CO_2)_{ca}}{p(H_2O)p(CO_2)_{an}}\right) \quad (13)$$

where E^{o} is the standard potential, the subscripts “an” and “ca” denote the anode and cathode, respectively, and p is the partial pressure of the gases. Other symbols have their usual meanings. E_{OCV} is a voltage at zero current, so it represents a theoretically maximum voltage in the cell. When the current flows in the cell, electrical resistance among the cell components and electrochemical resistance at the electrodes reduce the cell voltage. Thus, the performance of fuel cells is determined by the voltage loss, which is the difference between the open circuit voltage (E_{OCV}) and the voltage (V) at a current load (Eq. 14)

$$V{=}E_{OCV} - \eta_{IR} - \eta_{an} - \eta_{ca} \quad (14)$$

where η_{IR} is the ohmic loss due to the electrical resistance, and η_{an} and η_{ca} are the overpotential at the anode and cathode electrodes, respectively.

The ohmic loss is relatively easy to understand because the electrical resistance of the cell components behaves as a cause of voltage loss. However, determination of overpotential from the electrochemical reaction resistance at the electrodes has been an interesting research topic. The fuel cell electrodes require a large surface area to increase the reaction rate, and thus porous materials are employed. In addition, the electrode surface is covered by thin electrolyte film to provide the three-phase boundary of gas-liquid-solid where the electrochemical reaction occurs. Thus, the electrochemical resistance in MCFC is comprised of charge-transfer resistance on the electrode surface and mass transfer through the liquid film and gas channel as shown in Fig. 2.

At the anode, the following overpotential relations have been suggested based on the electrode kinetics of Eq. 8 by Selman’s group in the USA (Eq. 15a) [31] and CRIEPI (Central Research Institute of Electric Power Industry) in Japan (Eq. 15b) [32].

$$\eta_{an} = a_1 \cdot p(H_2)^{-0.42} p(CO_2)^{-0.17} p(H_2O)^{-1.0} \cdot i \quad (15a)$$

$$\eta_{an} = a_2 \cdot p(H_2)^{-0.5} \cdot i \quad (15b)$$

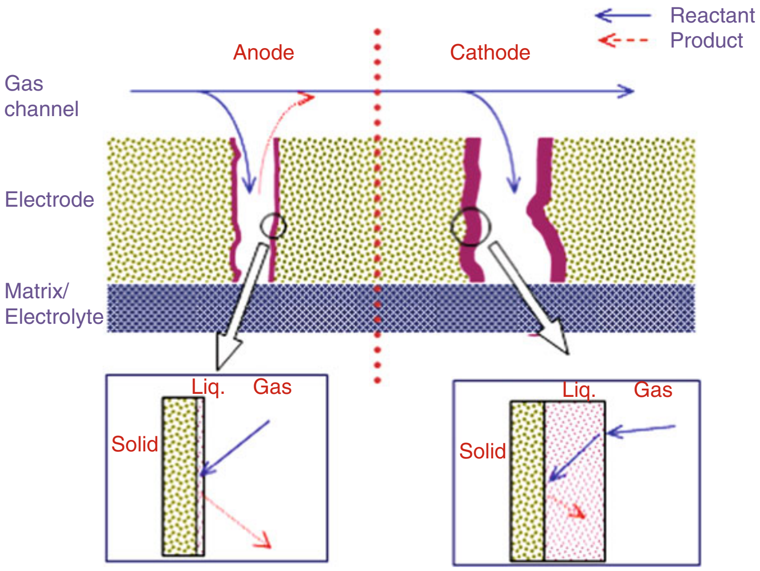
where a is the constant, and i is the current. Both relations were obtained by the steady-state polarization method. On the other hand, the overpotential relations at the cathode have been reported by Selman’s group (Eq. 16a) [31] and the CRIEPI group (Eq. 16b) [33].

$$\eta_{ca} = a_3 \cdot p(O_2)^{-0.43} p(CO_2)^{-0.09} \cdot i \quad (16a)$$

$$\eta_{ca} = \left(a_4 \cdot p(O_2)^{-0.75} p(CO_2)^{0.5} + a_5 \cdot p(CO_2)^{-1}\right) \cdot i \quad (16b)$$

Equation 16a is an empirical relation based on the electrode kinetics of a superoxide path (Eq. 9d). On the other hand, Eq. 16b resulted from the assumptions of mass-transfer resistance of superoxide ions and CO_2 in the carbonate electrolyte film.

Most of the electrode kinetics in molten carbonates has been investigated with half-cell experiments that used smooth surface electrodes. Various experimental techniques could be applied such as the use of rotating disk electrode (RDE) [34], rotating wire electrode [35], ultramicroelectrode [36], potential step [37], AC impedance [37], coulostatic relaxation [37],



Molten Carbonate Fuel Cells. Figure 2
Schematic drawing of reaction characteristics at the anode and cathode of MCFC

voltammetry [17], and so on. However, the MCFC uses a porous electrode in the cell. Since it is covered by a thin electrolyte film and has a pore volume of over 50%, the behavior of the porous electrode in the cell would be significantly different from that of a plain surface electrode in carbonate melts. A limited number of experimental methods could be applied for the investigation of the reaction characteristics of MCFCs. In general, the voltage loss due to ohmic loss and reaction overpotential in the MCFC has been analyzed with steady-state polarization [31], current interruption [38], and AC impedance methods [39]. As relatively new investigation tools, inert gas step addition (ISA) [22] and reactant gas addition (RA) [40] methods are introduced and the relationship between the methods is treated in this work.

Steady-State Polarization

Steady-state polarization (SSP) is a very simple method that measures voltage by applying currents with sufficient time intervals. Figure 3 shows some results according to different utilizations. An E_{OCV} of 1.07 V is observed. It is very close to the theoretical value according to Eq. 13 because the inlet composition of anode gas is H_2:CO_2:H_2O = 0.69:0.17:0.14 atm and that of cathode gas is air:CO_2 = 0.7:0.3 atm.

Since the utilization (u) is a ratio of consumed gas amounts to supplied amounts, it indicates a gas flow rate at a fixed current density.

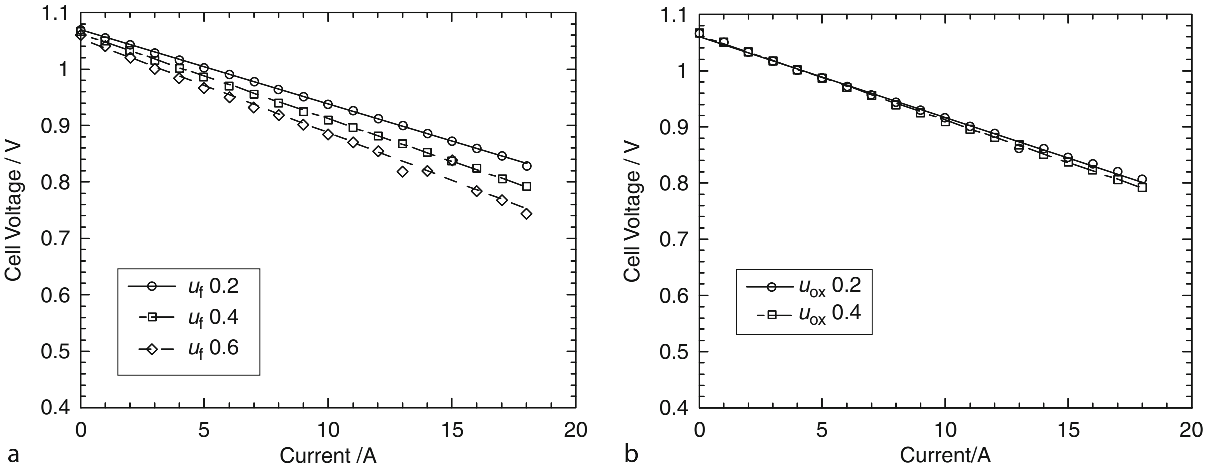
$$u = \frac{\text{consumed gas amounts}}{\text{supplied gas amounts}} \qquad (17)$$

The utilizations in the figures are based on currents of 15 A. The linear current-voltage behaviors are observed at the utilizations. This indicates that the MCFC has very low charge-transfer resistance as mentioned in the introduction. The difference between E_{OCV} and V at a current load is the total voltage loss according to Eq. 14. As shown in the figures, SSP cannot distinguish between the voltage losses accounted for by η_{IR}, η_{an}, and η_{ca}.

Figure 3 also shows that cell voltage is more severely dependent on the anode utilization than on the cathode one. This indicates that anodic overpotential is more affected by the flow rate than the cathodic one. It is a specific feature of MCFC that flow rate affects cell voltage.

Current Interruption

Current interruption (C/I) is a voltage relaxation method. The measurement is quite simple: the applied currents are rapidly interrupted and the following



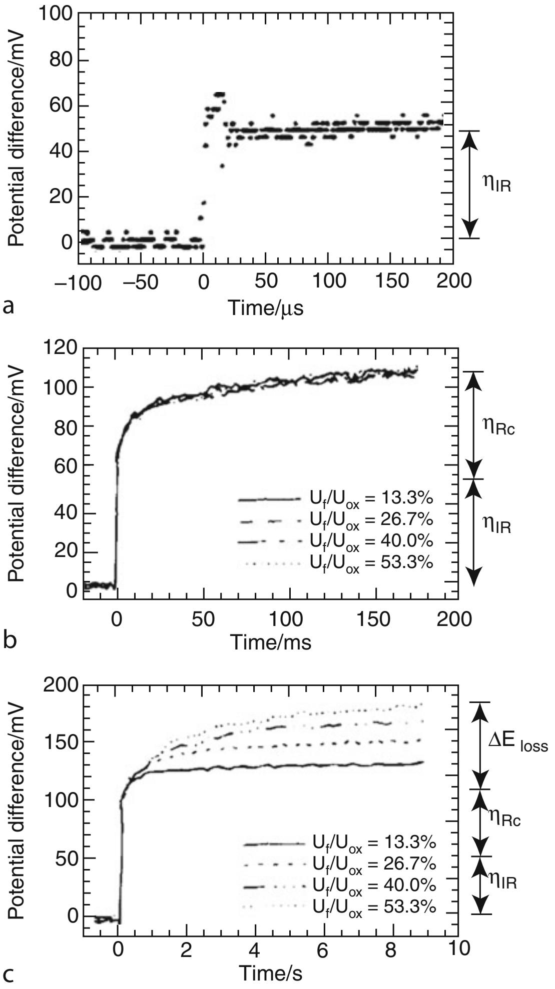
Molten Carbonate Fuel Cells. Figure 3

Steady-state polarization results at various utilizations (15 A current bases) with a 100 cm^2 class MCFC single cell at 650°C, 1 atm. (**a**) anode utilizations at a fixed cathode utilization of 0.4; (**b**) cathode utilizations at a fixed anode utilization of 0.4

voltage relaxation is recorded. Figure 4 presents some results of C/I measurement. In particular, Fig. 4a shows voltage relaxation during 200 μs. At the exact moment of interruption, a voltage jump due to the relaxation of electrical resistance occurs. In the MCFC, most of the electrical resistance is attributed to the ionic resistance in the electrolyte, and thus the voltage jump shows that the ionic resistance is relaxed right after the interruption. Then the flow rate independent time region follows as shown in Fig. 4b. However, voltage relaxation in the time region depends on the oxidant gas composition [38]. This implies that the time region represents the mass-transfer effect through the liquid electrolyte at the cathode. Then the longest time region of 10 s shows that voltage relaxation depends on the anode utilization; higher utilization requires a longer relaxation time. CRIEPI reported that the voltage relaxation for several seconds was ascribed to the relaxation of concentration distribution in the anode electrolyte film [41]. Consequently, the C/I method was found to show ohmic loss and anodic and cathodic overpotential for different time ranges.

AC Impedance

The AC impedance method is a powerful technique for electrode kinetics and mass-transfer investigation. The charge-transfer and mass-transfer resistances are an electrically parallel circuit with an electrochemical double layer that behaves as a capacitor. The parallel circuit of a resistance and a capacitor at a smooth surface electrode has the characteristic behavior of an AC signal: a 90° phase angle between the current and voltage signal at a high frequency AC signal and 0° at a low frequency signal. The phase angle is generally represented in the complex plane where the X axis represents the resistance component and the Y axis the capacitive one. Thus, the impedance of the parallel circuit draws a half circle due to the frequency change, and the diameter of the circle represents the resistance value of the circuit. Interpretation of the AC impedance in the MCFC has not been in agreement, although a lot of theoretical interpretation has been attempted. Difficulties in the interpretation of porous electrode behaviors are major reasons. Figures 5a and b show the results of AC impedance with different flow rates at the anode and cathode, respectively. They were measured in an open-circuit state, and thus electrodes were maintained in an equilibrium state. The length of the X axis from 0 to the high frequency initial point represents the internal resistance of the cell. The high frequency semicircle on the left has a frequency range of 1 kHz–5 Hz, which does not depend on the anode and cathode flow rate. In a previous work, it was reported that the high



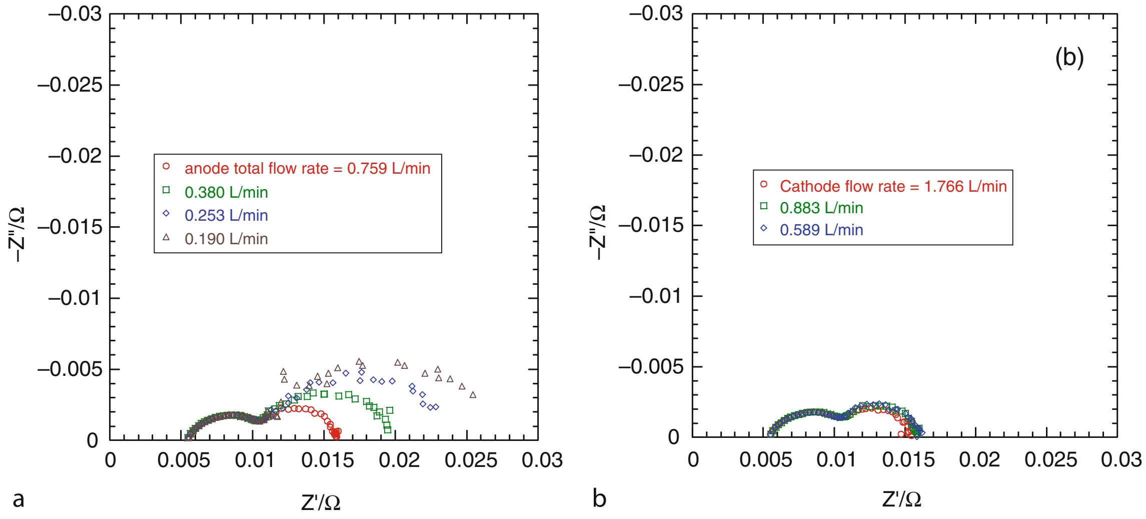
Molten Carbonate Fuel Cells. Figure 4
Voltage relaxations after current interruption of 10 A current loads at a 100 cm^2 class MCFC single cell (From [38]). (**a**) 200 μs time range; (**b**) 200 ms; (**c**) 10 s

frequency semicircle reflected cathodic overpotential due to the mass-transfer resistance through the liquid electrolyte at the cathode [42]. The semicircle on the right has a frequency range of 1 to 0.01 Hz, and is called the low frequency semicircle (LFSC). The LFSC shows a clear dependence on the anode gas flow rate and insignificant change due to the cathode flow rate as shown in the figure. This is in agreement with a previous work [42]. The enlarged LFSC at low flow rate indicates higher resistance in the cell. The CRIEPI group reported that the LFSC represents the effect of gas flow on the concentration distribution along the gas flow path on the electrode; a higher flow rate provides less concentration distribution [43].

Inert Gas Step Addition (ISA)

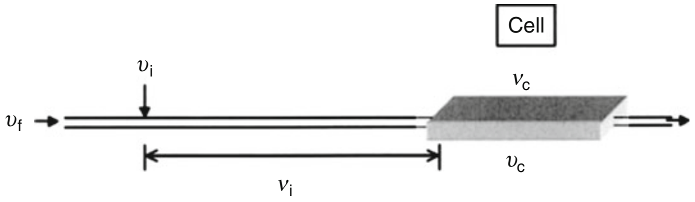
The above relations of anode overpotential (Eqs. 15a and 15b) and cathode overpotential (Eqs. 16a and 16b) only employ the partial pressure effect of gases. A lot of previous works on the electrode kinetics have been done with a coin type single cell, which had a geometric electrode area of about 3 cm^2. They were carried out with a very low gas utilization that was sufficient to neglect the gas-phase mass-transfer effect. The low utilization, however, was far from the actual condition where the anode utilization is normally over 70%. In addition, the performance of MCFC depends on the gas flow rate even in the 100 cm^2 class single cell as shown in Fig. 3. To investigate the flow rate effect in the MCFC, the inert gas step addition (ISA) method has been developed [22]. ISA can vary the utilization without changing gas compositions, and thus the flow rate effect could be analyzed.

Measurements ISA measurement was mainly carried out with 100 cm^2 class single cells because it had sufficient anode and cathode gas volumes to show the gas flow effect at the electrodes. In fact, the reactant gases flow through the gas channel over the electrode, and thus an electrode has a certain gas volume between the electrode and cell frame. Figure 6 shows the gas flow path of an electrode. When the gas line volume, v_i, is bigger than the gas volume of an electrode, v_c, the added inert gas enlarges the reactant gas flow rate in the electrode during the time range of $t_{i,a}$ of Fig. 7. The flow rates of reactant gases are enhanced without partial pressure change during $t_{i,a}$, which results in a utilization shift. Thus, the voltage variation during the time range represents the utilization effect on the overpotential. The added inert gas flows inside the cell until the step off of inert gas, which varies the gas partial pressures between $t_{i,a}$ and $t_{i,b}$. Therefore, ISA also provides a partial pressure effect on the overpotential. When the inert gas flow is interrupted, the remaining inert gas in the volume, v_i, flows in the cell during $t_{i,b}$. Thus the reactant flow rate is decreased during the time range. The flow rate of inert gas was controlled with a mass flow controller (MFC). During



Molten Carbonate Fuel Cells. Figure 5

AC impedance results with various gas flow rates with a 100 cm^2 class MCFC single cell at 650°C, 1 atm, OCV, 5 mV rms signal, 1 kHz–0.01 Hz. (**a**) Various anode flow rates at a fixed cathode flow rate of 0.883 L min^{-1}; (**b**) Various cathode flow rates at a fixed anode flow rate of 0.759 L min^{-1}



Molten Carbonate Fuel Cells. Figure 6

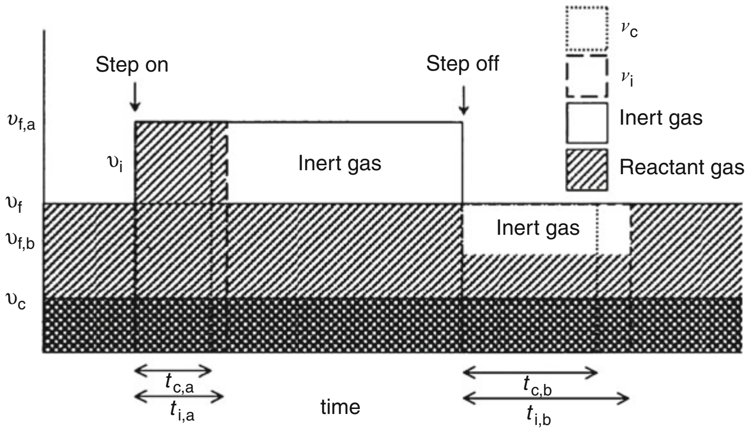
Schematic drawings of gas flow path of an electrode (From [22]). υ_f is the reactant flow rate, υ_i is the inert gas flow rate, υ_c is the consumed gas rate by the applied currents, v_i is the volume between inert gas inlet port and electrode, v_c is the volume of gas channel

the flow change, the cell voltage and inert gas flow rate were simultaneously recorded with an oscilloscope.

Analysis of Cell Behavior As shown in Fig. 2, MCFC reactions are comprised of charge-transfer reactions on the electrode surface and mass-transfer processes through the gas and liquid phases in series. Resistances in those processes are reciprocal numbers of reaction-rate constant and mass-transfer coefficients that are represented in the overpotential. Thus the overpotential relation is as follows [22]:

$$\eta = \frac{iRT}{n^2F^2ap_0}\left(\frac{h}{k_0}+\frac{h}{k_L}+\frac{1}{k_G}\right) \qquad (18)$$

where i is the current, a is the geometrical area, h is the Henry's Law constant, p_0 is the bulk gas pressure, k_0 is the reaction-rate constant, and k_L and k_G are the mass-transfer coefficients through the liquid electrolyte and gas phase, respectively. Other symbols have their usual meanings. As mentioned in the chapter of anode electrode, the H_2 oxidation rate is sufficiently fast ($i_o \approx 100$ mA cm^{-2}) for the charge-transfer resistance to be neglected. In addition, the electrolyte film on the anode electrode can be assumed to be negligibly thin according to the dry agglomerate model [3], and then the mass-transfer resistance through the electrolyte film can be neglected. Consequently, the anode is assumed to be a gas-phase mass-transfer control



Molten Carbonate Fuel Cells. Figure 7
Schematic drawings of flow rate changes caused by the inert gas addition. The reactant flow rate increases during $t_{i,a}$ and decreases for $t_{i,b}$ (From [22]). υ_f is the reactant flow rate, υ_i is the inert gas flow rate, υ_c is the consumed gas rate due to the applied currents, $\upsilon_{f,a}$ is the flow rate increase due to the inert gas addition, $\upsilon_{f,b}$ is the flow rate decrease by the interruption of inert gas addition, v_i is the volume between inert gas inlet port and electrode, v_c is the volume of gas channel. t_c is the time to fill v_c, t_i is the time of the gas filled in the volume v_i to flow over the electrode

process. The gas-phase mass-transfer coefficient (k_G) was obtained from the mass-transfer coefficient (k_B) of the boundary layer theory in the case of mass transfer between laminar flow and plain substrates [22]. Then k_G is expressed as follows:

$$k_B \cong 0.664\left(\frac{\upsilon_f}{L}\right)^{\frac{1}{2}}(D_G)^{\frac{2}{3}}(\nu)^{-\frac{1}{6}} \tag{19}$$

For the unit conversion,

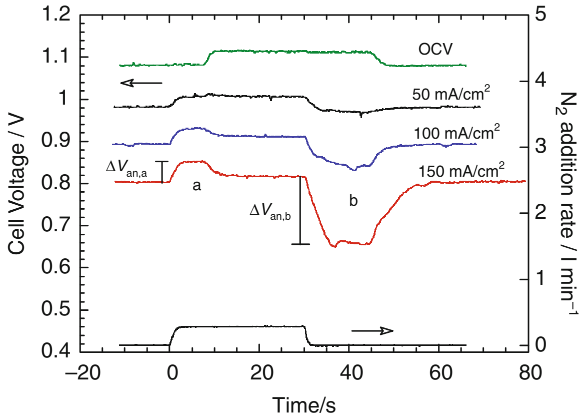
$$k_G = \frac{k_B}{RT} \tag{20}$$

where υ_f is the flow velocity, L is the electrode length, D_G is the gas diffusivity, and ν is the kinematic viscosity. Other symbols have their usual meanings. Since the gas velocity corresponds to gas flow rate, Eq. (20) can be expressed in terms of utilization (u). Then Eq. (18) becomes:

$$\eta_{an} \cong \eta_G = i\frac{RT}{n^2F^2ap_0}\left(\frac{1}{k_G}\right) = q \cdot u^{0.5} \tag{21}$$

where $q = 1.51\frac{R^2T^2(iLs)^{1/2}\nu^{1/6}}{(n^3F^3a^2p_0)^{1/2}D_G^{2/3}}$, and s is the cross section area of the gas channel.

Figure 8 shows voltage behaviors with 0.3 L min^{-1} N_2 addition to the anode at various current densities. At the open-circuit state, the voltage increases due to the addition because E_{OCV} rises with decreasing partial pressure of anode gases according to Eq. 13. However, at polarization states, a positive peak, “a,” and a negative voltage peak, “b,” are observed. The peaks are due to the change in the flow rate of reactant gases as shown in Fig. 7. The flow rate increase results in the positive “a” peak while the flow rate decrease leads to the negative



Molten Carbonate Fuel Cells. Figure 8
Voltage shift patterns with an addition of 0.3 L min^{-1} N_2 to the anode at 650°C, u_f = 0.6 (0.253 L min^{-1}) and u_{ox} = 0.4 (0.883 L min^{-1}) of i = 150 mA cm^{-2} (From [47])

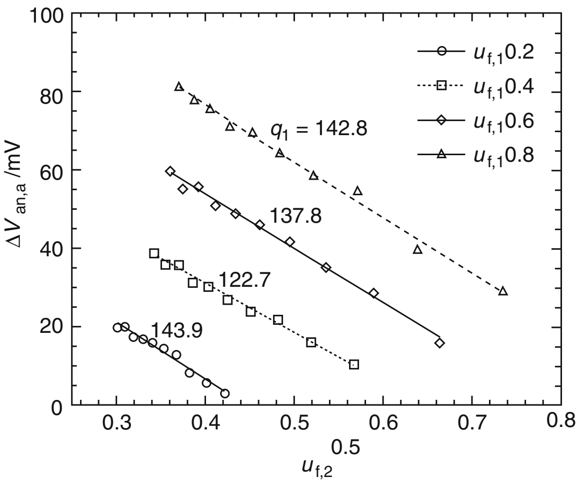
"b" peak. This indicates that the anode reaction is affected by the flow rate of anode reactant gases.

Substituting N_2 with Ar and He gives identical voltage behaviors [22]. This shows that ISA measurements are available regardless of the inert gas species. However, the voltage between the "a" and "b" peaks was affected by the inert gas species; helium showed the highest voltage among them [22]. During the time range, inert gas flows inside the cell and the gas species may affect the diffusivity of H_2. Indeed, helium has the highest diffusivity among them, and thus helium provided the lowest mass-transfer resistance in the anode. This strongly implies that the anode reaction is a gas-phase mass-transfer control process.

Since the height of the positive peak, $\Delta V_{\mathrm{an,a}}$, is an overpotential difference at the flow rate change and the flow rate corresponds to the utilization, the peak height can be expressed in terms of anode utilization (u_{f}).

$$\Delta V_{an,a} = \eta_{an1} - \eta_{an2} = qu_{\mathrm{f},1}^{0.5} - qu_{\mathrm{f},2}^{0.5} = m - qu_{\mathrm{f},2}^{0.5} \quad (22)$$

where subscripts 1 and 2 denote before and after the N_2 addition respectively and m is a constant. Thus the peak height has a linear relation with utilization. Figure 9 shows the peak heights with various N_2 addition rates at different anode utilizations and arranges the heights according to the relation of Eq. 22. The linearity of Eq. 22 at the different anode gas utilizations verifies the validity of the equation, which shows that the anode reaction is a strong gas-phase mass-transfer control process. The slope of q is about 140 mV and the anodic overpotential can be obtained according to Eq. 21.



Molten Carbonate Fuel Cells. Figure 9
Rearrangement of the peak height in Fig. 8 with shifted utilization according to Eq. 22 (From [22])

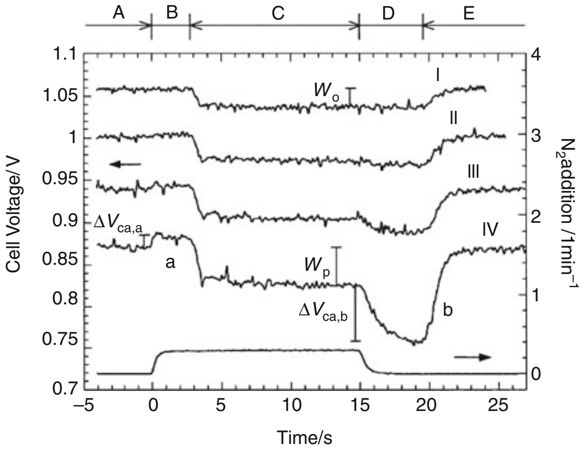
On the other hand, the linear current-voltage behavior of Fig. 3 and the very high reaction rate of the cathode ($i_{\mathrm{o}} \approx 10$ mA cm^{-2}) allow the assumption of negligible charge-transfer resistance at the cathode. In fact, the cathode has a relatively thick carbonate electrolyte on the surface as shown in Fig. 2. Thus, the following relation has been suggested as the cathodic overpotential [22].

$$\eta_{\mathrm{ca}} \cong \eta_{\mathrm{ca,G}} + \eta_{\mathrm{ca,L}} = i\frac{RT}{n^2F^2ap_0}\left(\frac{h}{k_{\mathrm{ca,L}}} + \frac{1}{k_{\mathrm{ca,G}}}\right) \quad (23)$$

Equation 23 is the sum of overpotential due to the liquid and gas phases, and thus their separate estimation is available. When the overpotential due to the gas-phase resistance is considered, the following relation can be used:

$$\eta_{\mathrm{ca,G}} = i\frac{RT}{n^2F^2ap_0}\left(\frac{1}{k_{\mathrm{ca,G}}}\right) = q \cdot u_{\mathrm{ox}}^{0.5} \quad (24)$$

The N_2 addition to the cathode brings about two positive and negative voltage peaks in the "B" and "D" time regions as shown in Fig. 10. The origins of the



Molten Carbonate Fuel Cells. Figure 10
Cathodic ISA results at different currents with nitrogen addition rate of 0.3 L min^{-1} at a cathode flow rate of 0.589 L min^{-1} and anode flow rate of 0.756 L min^{-1}, 923 K, 1 atm, I/0; II 50; III 100; IV 150 mA cm^{-2} (From [22])

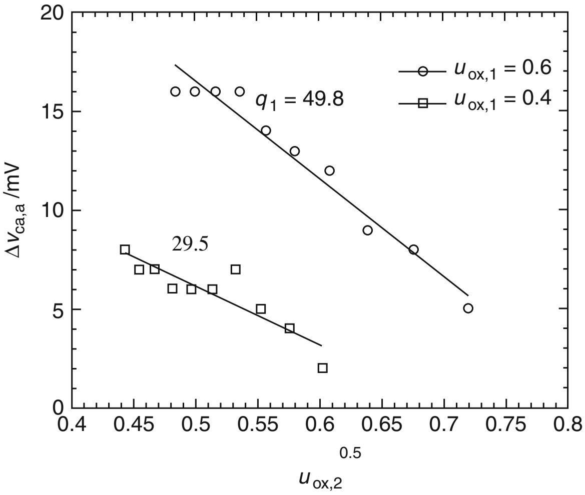
peaks are due to the flow rate and utilization change similar to the anode as shown in Fig. 8. Applying the relation of Eq. 22 to the cathode, we obtain Eq. 25.

$$\Delta V_{\mathrm{ca,a}} = \eta_{\mathrm{ca1}} - \eta_{\mathrm{ca2}} = q u_{\mathrm{ox,1}}^{0.5} - q u_{\mathrm{ox,2}}^{0.5} = m - q u_{\mathrm{ox,2}}^{0.5} \tag{25}$$

Then the q value of the cathode represents overpotential due to the gas-phase mass transfer at the cathode.

Figure 11 shows the results of Eq. 25 at the cathode. The value, q_1, at the cathode equal to q is very small compared with that at the anode of Fig. 9, although the small positive peak height provides deviations in the $\Delta V_{\mathrm{ca,a}}$. This means that the cathode has much smaller overpotential due to the gas transport in the cell. In addition, q_1 values are different with cathode utilization; they have larger values at a higher utilization. In general, the cathode showed higher overpotential of over 50% of oxidant utilization [22]. The low diffusivity in the gas phase and low O_2 solubility in the carbonate melts could be the reason. The q_1 value indicates that the cathodic overpotential also depends on the utilization.

On the other hand, the voltage in the "C" time region of Fig. 10 is a steady-state value of voltage at the N_2 flowing in the cathode. The N_2 in the cell varies the gas partial pressures, and thus the region represents some effects of gas partial pressures on the overpotential. Adding the same amount of N_2 to the various gas partial pressures of O_2:CO_2:N_2 (Table 3) gives overpotential differences ($\Delta W = W_P – W_O$) related to the gas partial pressure. In this case, W_P is the voltage gap at a polarization state and W_O is that at open-circuit state by N_2 addition to the cathode as shown in Fig. 10. Thus, ΔW has the following relations:



Molten Carbonate Fuel Cells. Figure 11 Rearrangement of the peak heights in Fig. 10 with shifted utilization according to Eq. 25 (From [22])

$$\Delta W = \eta_{\mathrm{ca,L,2}} - \eta_{\mathrm{ca,L,1}} = \left(R_{\mathrm{ca,L2}} - R_{\mathrm{ca,L1}}\right) \cdot i/a \tag{26}$$

where subscripts 1 and 2 represent before and after N_2 addition, respectively, and a is the geometrical electrode area. Then ΔW has a linear relation with gas partial pressure with Eq. (11b).

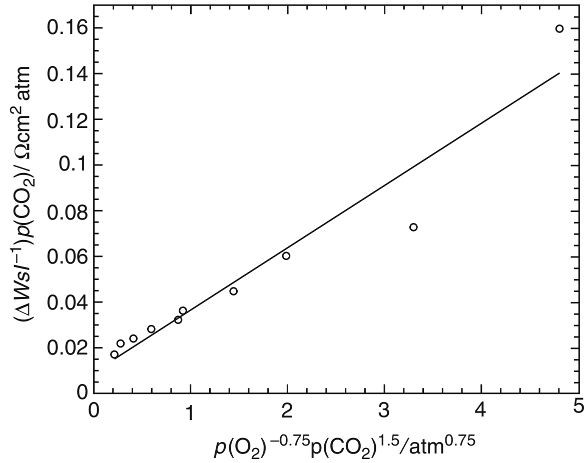
$$\begin{aligned}(\Delta W \cdot a/i)p(CO_2) &= \left(R_{\mathrm{ca,L2}} - R_{\mathrm{ca,L1}}\right)p(CO_2)\\ &= A' p(O_2)^{-0.75} p(CO_2)^{1.5} + B'\end{aligned} \tag{27}$$

where $A' = A(\beta^{-0.25} - 1), B' = B(\beta^{-1} - 1), \quad \beta = p(O_2)_2/p(O_2)_1$, A and B are the constants of Eq. 11b, and the subscripts 1 and 2 represent before and after N_2 addition, respectively.

When we assume that the cathode overpotential due to the mass transfer through the carbonate electrolyte is combined diffusion control of superoxide ions and CO_2, the overpotential is a function of gas partial pressure as shown in Eqs. 11a and 11b. Equation 11b shows a linear relation between the ΔW and gas partial pressures. Figure 12 shows linearity of Eq. 11b, indicating that the mass-transfer resistance through the electrolyte film causes cathodic overpotential. From Eq. 26 we can obtain A and B values. Then with Eq. 11b we can have $R_{\mathrm{ca,L}}$ and $\eta_{\mathrm{ca,L}}$ at normal gas partial pressures of $p(O_2) = 0.15$ atm and $p(CO_2) = 0.3$ atm. The value of $\eta_{\mathrm{ca,L}}$ under this

Molten Carbonate Fuel Cells. Table 3 Various compositions of the cathode gases

Gases	Ratio (atm)
O_2:CO_2:N_2	0.9:0.1:0,
	0.7:0.3:0, 0.7:0.1:0.2,
	0.5:0.5:0, 0.5:0.3:0.2, 0.5:0.1:0.4,
	0.3:0.7:0, 0.3:0.5:0.2, 0.3:0.3:0.4,



Molten Carbonate Fuel Cells. Figure 12
Relationships of the cathodic overpotentials with partial pressures of oxygen and CO_2 according to Eq. 27 at 923 K, 1 atm (From [22])

condition is about 62 mV, which is much larger than $\eta_{ca,G}$ ($\approx$18 mV at $u_{ox} = 0.4$) from Eq. 24. This means that overpotential at the electrolyte film is much larger than that at the gas phase and the cathodic reaction is mostly the liquid-phase mass-transfer control process.

The above results show that the anodic overpotential is mostly attributed to the gas-phase mass-transfer resistance and the cathodic one is a sum of overpotential due to gas-phase resistance ($\eta_{ca,G}$) and liquid-phase resistance ($\eta_{ca,L}$). Thus, the following relation can be suggested:

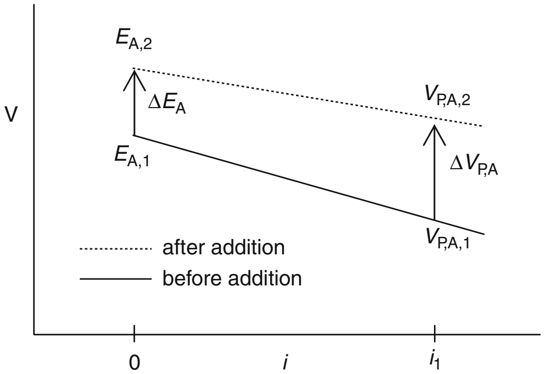
$$\eta_{an} \cong \eta_{an,G} = q_{an} \cdot u_f^{0.5} \tag{28a}$$

$$\eta_{ca} \cong \eta_{ca,G} + \eta_{ca,L} = q_{ca} \cdot u_{ox}^{0.5} + A'' \cdot p(O_2)^{-0.75} p(CO_2)^{0.5} + B'' \cdot p(CO_2)^{-1} \tag{28b}$$

where q_{an} and q_{ca} are the constants q of Eq. 22 and 25, respectively, $A'' = A \cdot i/a$ and $B'' = B \cdot i/a$, and A and B are the constants of Eq. 27.

Reactant Gas Addition (RA) Method

The cathodic overpotential from the ISA method, the sum of $\eta_{ca,L}$ and $\eta_{ca,G}$, is only 80 mV at $u_{ox} = 0.4$. This is smaller than the anodic overpotential from the method at $u_f = 0.4$, which is about 90 mV according to Eq. 21. This is contradictory to the conventional concept that the cathodic overpotential is larger than the anodic one because of the slow oxygen reduction at the cathode.



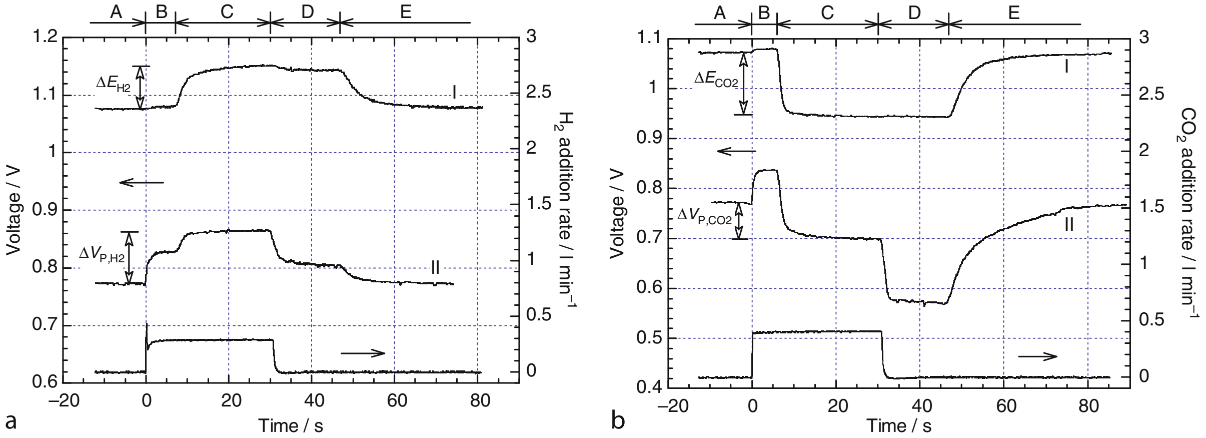
Molten Carbonate Fuel Cells. Figure 13
Schematic drawings of voltage shift behaviors by the addition of a reactant gas at open-circuit (ΔE_A) and polarization states ($\Delta V_{P,A}$) (From [41])

Meanwhile, the anode and cathode reactions of Eqs. 1a and 1b are multi-component reaction systems. As mentioned regarding the ISA method, the anode and cathode reactions are mass-transfer control processes. Then the mass-transfer of each species would provide overpotential due to the species. To investigate the overpotential attributed to each species, the reaction gas addition method was attempted. This is very similar to the ISA except that a reactant gas is added to an electrode instead of an inert gas [41]. Figure 13 shows the voltage behaviors due to the addition of a reactant gas. Here, the subscript A denotes a reactant gas species.

When a certain amount of a reactant gas is added to an electrode at the open-circuit state, the addition changes the partial pressures of the cell and determines E_{OCV} according to Eq. 13. When the same amounts of reactant gas are added at a polarization state, the voltage is varied by overpotential according to Eq. 14. Thus the gap (ΔV_A) between the voltage shift at the open-circuit state (ΔE_A) and at a polarization state ($\Delta V_{P,A}$) is overpotential variation due to the addition.

$$\Delta V_A = \Delta V_{P,A} - \Delta E_A \tag{29}$$

where $\Delta V_{P,A} = V_{P,2} - V_{P,1}$ and $\Delta E_A = E_{OCV,2} - E_{OCV,1}$. When $\Delta V_A > 0$, the addition mitigates mass-transfer resistance at the electrode. On the

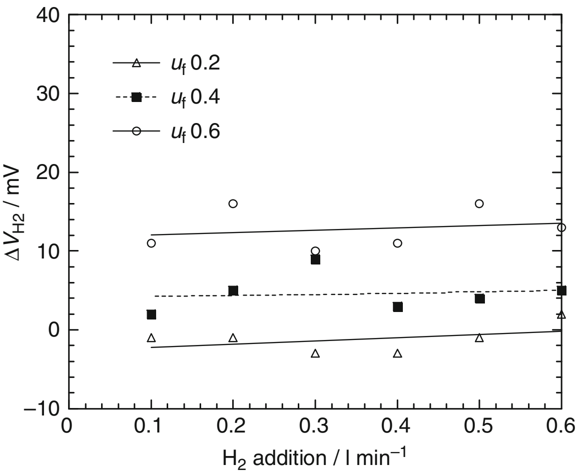


Molten Carbonate Fuel Cells. Figure 14

Results of the RA measurement at the anode with conditions of 923 K, 1 atm at open-circuit state (I) and polarization state of 150 mA cm^{-2} (II), anode feed rate = 0.253 L min^{-1} (u_f = 0.6 at 150 mA cm^{-2}), and cathode feed rate = 0.883 L min^{-1} (u_{ox} = 0.4 at 150 mA cm^{-2}) (From [41]). (**a**) 0.3 L min^{-1} H_2 addition; (**b**) 0.4 L min^{-1} CO_2 addition

contrary, $\Delta V_A < 0$ indicates that the addition enlarges the resistance. When $\Delta V_A = 0$, the addition does not affect the resistance.

Figure 14a shows RA results with H_2 addition to the anode. At the open-circuit state (curve I) the added H_2 flows inside the cell in the "C" and "D" time regions, which results in a partial pressure and E_{OCV} increase in these regions. At a polarization state of 150 mA cm^{-2} (curve II) two positive voltage steps in the "B" and "C" time regions are observed. The step in the "B" region is due to the increase in flow rate of reactant gases due to the H_2 addition, which is the same as the reason for the positive voltage peak obtained with the ISA method. For the "C" region the enlarged H_2 flow rate results in the second voltage step, which involves a change in overpotential due to the addition. The voltage shift of curve II ($\Delta V_{P,H2}$) is larger than that at the open-circuit state (ΔE_{OCV}). In principle, the gas inlet conditions for the two curves are identical. Thus, the difference (ΔV_{H2}) represents variation in overpotential due to the H_2 addition. This results in a positive ΔV_{H2} according to Eq. 29, which shows that H_2 addition reduces the anodic overpotential. It also implies that the anode has overpotential due to the mass-transfer resistance of H_2 species. However, the behavior can provide information on overpotential due to H_2 species through ΔV_{H2}. When the H_2 addition rate was varied from 0.1 to 0.6 L min^{-1}, ΔV_{H2} showed consistency in the rates as shown in Fig. 15. This means that the resistance due to the H_2 species is sufficiently reduced by the addition, and then a consistent ΔV_{H2} is obtained. In addition, when anodic utilization, u_f, is increased, ΔV_{H2} is enlarged [41]. This also implies that anodic overpotential depends on the H_2 flow rate; a lower H_2 flow rate has a larger anodic overpotential.

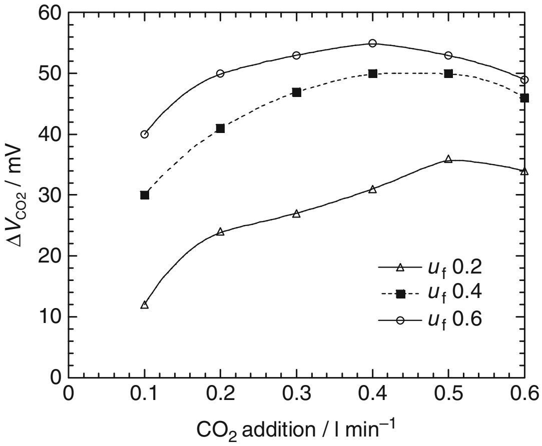


Molten Carbonate Fuel Cells. Figure 15

ΔV_{H2} behavior with respect to the H_2 addition amount at the various anode utilizations at 923 K, 1 atm, cathode feed rate = 0.883 L min^{-1} (From [41])

Figure 14b shows the results of CO_2 addition to the anode. Indeed, CO_2 is a product species of the anode reaction as shown in Eq. 1a. Thus the CO_2 addition reduces E_{OCV} according to Eq. 13. At open-circuit state the reduced E_{OCV} is observed due to the CO_2 addition in the "C" and "D" time regions. At a current density of 150 mA cm^{-2}, two voltage peaks are observed; like the voltage peaks obtained using the ISA method ascribed to the change in flow rate of anode gas. The voltage shift due to the addition at the current density ($\Delta V_{P,CO2}$) is much smaller than that at the open-circuit state (ΔE_{CO2}). Since those values are negative, the difference (ΔV_{CO2}) is a positive value. This indicates that CO_2 addition to the anode reduces anodic overpotential. This can be explained by the reaction kinetics of Eq. 8, where the reaction rate has a positive order for the CO_2 species. Therefore raising the CO_2 partial pressure reduces anodic overpotential [41]. Analysis of overpotential with various anode gas compositions shows identical overpotential behavior, whereby increasing CO_2 partial pressure reduces anodic overpotential [44].

Dissimilar to the H_2 addition, ΔV_{CO2} depends on the amount of CO_2 addition (Fig. 16). Furthermore, ΔV_{CO2} values are much larger than ΔV_{H2}, which implies that CO_2 species has higher mass-transfer resistance than H_2 species and anodic overpotential is more dependent on the CO_2 flow rate. The CO_2 addition reduces the resistance, and thus ΔV_{CO2} rises with CO_2 addition amounts as shown in Fig. 16. At certain amounts, from 0.4 to 0.5 L min^{-1}, ΔV_{CO2} has a maximum value. This indicates that the mass-transfer resistance due to the CO_2 species becomes a minimum. Over the amounts, the CO_2 species rather signifies the resistance probably due to reducing the mass transfer of H_2 species. Then ΔV_{CO2} decreases again. ΔV_{CO2} also depends on the anodic utilization; higher utilization shows larger overpotential. Similar to the H_2 addition, this indicates that the anode reaction is a mass-transfer control process of CO_2.

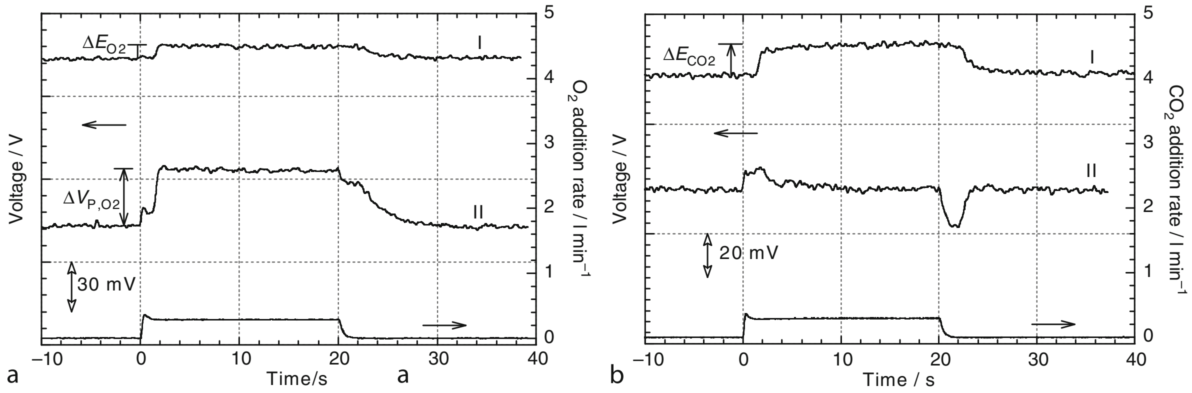


Molten Carbonate Fuel Cells. Figure 16
ΔV_{CO2} behavior with respect to the CO_2 addition amount at various utilizations of anode gas at 923 K, 1 atm, cathode feed rate = 0.883 L min^{-1} (From [41])

The H_2O addition also decreased anodic overpotential [41]. The anode gas was humidified with a bubbler that contained water at a certain temperature. Then the partial pressure of H_2O was controlled by the water temperature in the bubbler. The anodic overpotential decreased monotonously with the increase in H_2O content [41], although the H_2O addition reduced E_{OCV} according to Eq. 13. In addition, ΔV_{H2O} rose with the anodic utilization, which was given larger overpotential by the low H_2O flow rate. The positive order of H_2O partial pressure in Eq. 8 is also the reason for the overpotential behavior.

The above results indicate that anode gases of H_2, CO_2, and H_2O provide overpotential due to their mass-transfer limitations. Moreover, the anodic overpotential rises with utilization. These results indicate that the anode reaction is a mass-transfer control process of the species and that the anodic overpotential is a sum of overpotentials due to the mass-transfer resistance of the species. Interestingly ΔV_{CO2} and ΔV_{H2O} are much larger than ΔV_{H2} under normal operating conditions. The low flow rate of CO_2 and H_2O under the condition (H_2:CO_2:H_2O = 0.69:0.17:0.14 atm) can be a reason [44].

Figure 17a shows RA results with O_2 addition to the cathode. At open-circuit state the O_2 addition slightly enhances E_{OCV}. However, at a current density of 150 mA cm^{-2}, a very high voltage shift ($\Delta V_{P,O2}$) is observed. The small step at 0 s originates from the increase in the flow of reactant gases due to the addition. As mentioned in the section on ISA, the cathode has very small mass-transfer resistance in the gas phase, and thus the height is rather small.



Molten Carbonate Fuel Cells. Figure 17

Results of RA measurement at the cathode with conditions of 923 K, 1 atm at open-circuit state (I) and 150 mA cm^{-2} polarization state (II), anode feed rate = 0.759 L min^{-1} (u_f = 0.2 at 150 mA cm^{-2}), and cathode feed rate = 0.883 L min^{-1} (u_{ox} = 0.4 at 150 mA cm^{-2}) (From [41]). (**a**) 0.3 L min^{-1} O_2 addition; (**b**) 0.3 L min^{-1} CO_2 addition

Therefore, we can expect a large difference ($\Delta V_{O2} = \Delta V_{P,O2} - \Delta E_{O2}$) between voltage shifts at open-circuit state (ΔE_{O2}) and at polarization state ($\Delta V_{P,O2}$). This shows that the cathode has significant overpotential due to the O_2 species and the cathode reaction is an oxygen mass-transfer limitation process. From the half-cell experiments it was also suggested that the mass-transfer limitation of O_2 species prevails at the cathode [45].

The CO_2 addition to the cathode enhances E_{OCV} according to Eq. 13 as shown in Fig. 17b. This is a very similar behavior to the O_2 addition. At a current density of 150 mA cm^{-2}, the CO_2 addition gives rise to the two voltage peaks. As mentioned in the section on the ISA method, these are ascribed to the change in cathode flow rate due to the addition; the positive one is due to the increase in flow rate and the negative one is due to the decrease in flow rate. A dominant feature is that the voltage difference at the current density, $\Delta V_{P,CO2}$, is almost zero. This means that ΔV_{CO2} is a negative value and the CO_2 addition to the cathode enlarges cathodic overpotential. Considering that CO_2 has about ten times the gas solubility of O_2 in the molten carbonate and that the mass-transfer resistance of O_2 species is dominant at the cathode, it is plausible that the CO_2 addition reduces O_2 partial pressure and enhances the resistance of O_2 species in the carbonate electrolyte. Consequently, the CO_2 species may not provide cathodic overpotential under normal operating conditions due to its high solubility.

Future Directions

MCFCs based on natural gas fuel have been commercialized across the world. In general, natural gas MCFCs are economically inferior to coal power electricity. More economic fuels such as decomposition gas of organic waste and coal gas, and so on, are required for wide scale use of MCFC. To investigate the validity of a new fuel, performance analysis tools should be prepared. The methods in this work can be utilized, but some improvements are also necessary. One is the establishment of a theoretical and experimental basis of the AC impedance method for the performance analysis of MCFCs. The strong point of convenient measurement of the method has been weakened by its obscure analytical basis. Another one is the verification of the relationships among the measurement tools listed in this work.

Another point to be mentioned for the dissemination of MCFCs is extending their lifetime. Electrolyte management is strongly related to the problem. The molten carbonate electrolyte is depleted mostly by corrosion with metals and weakening of electrolyte holding in the matrices. Cell design and surface treatment of metal should be considered. It is also necessary to search for appropriate material for the matrix.

Combinations of fuel cells also provide high efficiency. SOFC has been developed as a power system that is comparable to MCFC. SOFC is comprised entirely of solid materials, and thus cracking due to thermal shock is a fundamental problem. The problem confines the SOFC to relatively small power systems, so far. In general, SOFC has higher operation temperatures than MCFC. Thus, series combination of SOFC and MCFC may enhance power generation efficiency.

Bibliography

1. Yuh C, Hilmi A, Farooque M, Leo T, Xu G (2009) Direct fuel cell materials experience. ECS Trans 17:637–654
2. Selman JR, Maru HC (1981) Physical chemistry and electrochemistry of alkali carbonate melts. In: Mamantov G, Braunstein J (eds) Advances in molten salt chemistry. Plenum Press, New York, pp 202–212
3. Yuh CY, Selman JR (1984) Polarization of the molten carbonate fuel cell anode and cathode. J Electrochem Soc 131:2062–2069
4. Hong SG, Selman JR (2004) Wetting characteristics of carbonate melts under MCFC operating conditions. J Electrochem Soc 151:A77–A84
5. Morita H, Komoda M, Mugikura Y, Izaki Y, Watanabe T, Masuda Y, Matsuyama (2002) Performance analysis of molten carbonate fuel cell using a Li/Na electrolyte. J Power Sources 112:509–518
6. Nishina T, Masuda Y, Uchida I (1993) Gas solubility and diffusivity of H_2, CO_2 and O_2 in molten alkali carbonates. In: Saboungi ML, Kojima H, Duruz J, Shores D (eds) Proceedings of the international symposium on molten salt chemistry and technology (The Electrochemical Society PV93-9), pp 424–435
7. Yuh CY, Farooque M, Maru H (1999) Advances in carbonate fuel cell matrix and electrolyte. In: Uchida I, Hemmes K, Lindbergh G, Shores DA, Selman JR (eds) Carbonate fuel cell technology (The Electrochemical Society PV99-20), pp.189–201
8. Fujita Y (2003) Durability. In: Vielstich W, Lamm A, Gasteiger HA (eds) Handbook of fuel cells fundamentals technology and applications. Wiley, New York, pp 969–982
9. Yuh C, Johnsen R, Farooque M, Maru H (1993) Carbonate fuel cell endurance : Hardware corrosion and electrolyte management status. In: Shores D, Maru H, Uchida I, Selman JR (eds) Carbonate fuel cell technology, (The Electrochemical Society PV93-3), pp 158–170
10. Fujita Y, Nishimura T, Hosokawa JI, Urushibata H, Sasaki A (1996) Degradation of materials in molten carbonate fuel cells with Li/Na electrolyte. In the 3 rd FCDIC fuel cell symposium proceedings (Fuel Cell Development Information Center, Japan), pp 151–155
11. Matsumoto K, Yuasa K, Nakagawa K (1999) Protection against localized corrosion of stainless steel below 843 K in molten lithium-sodium carbonate. Denki Kagaku 67:253–258
12. Hoffmann J, Yuh CY, Jopek AG (2003) Electrolyte and material challenges. In: Vielstich W, Lamm A, Gasteiger HA (eds) Handbook of fuel cells fundamentals technology and applications. Wiley, New York, pp 921–941
13. Ang PGP, Sammells AF (1980) Influence of electrolyte composition on electrode kinetics in the molten carbonate fuel cell. J Electrochem Soc 127:1287–1293
14. Jewulski J, Suski L (1984) Model of isotropic anode in the molten carbonate fuel cell. J Appl Electrochem 14: 135–143
15. Lu SH, Selman JR (1984) Electrode kinetics of fuel oxidation at copper in molten carbonates. J Electrochem Soc 131: 2827–2833
16. Nishina T, Takahashi M, Uchida I (1990) Gas electrode reactions in molten carbonate media IV. Electrode kinetics and mechanism of hydrogen oxidation in (Li + K)CO_3 eutectics. J Electrochem Soc 137:1112–1121
17. Appleby AJ, Nicholson SB (1977) Reduction of oxygen in alkali carbonate melts. J Electroanal Chem 83:309–328
18. Appleby AJ, Nicholson S (1974) The reduction of oxygen in molten lithium carbonate. Electroanal Chem Interfacial Electrochem 53:105–119
19. Kinoshita K (1992) Electrochemical oxygen technology. Wiley, New York, p 37
20. Nishina T, Uchida I, Selman JR (1994) Gas electrode reactions in molten carbonate media V. Electrochemical analysis of the oxygen reduction mechanism at a fully immersed gold electrode. J Electrochem Soc 141:1191–1198
21. Yoshikawa M, Mugikura Y, Watanabe T, Ota T, Suzuki A (1999) The behavior of MCFCs using Li/K and Li/Na carbonates as the electrolyte at high pressure. J Electrochem Soc 146: 2834–2840
22. Lee CG, Kang BS, Seo HK, Lim HC (2003) Effect of gas-phase transport in molten carbonate fuel cell. J Electroanal Chem 540:169–188
23. Uchida I, Mugikura Y, Nishina T, Itaya K (1986) Gas electrode reactions in molten carbonate media II. Oxygen reduction kinetics on conductive oxide electrodes in (Li + K)CO_3 eutectic at 650 °C. J Electroanal Chem 206:241–252
24. Baumgartner C (1984) Electronic conductivity decrease in porous NiO cathodes during operation in molten carbonate fuel cell. J Electrochem Soc 131:2607–2610
25. Ota K, Mitsushima S, Kato S, Asano S, Yoshitake H, Kamiya N (1992) Solubilities of nickel oxide in molten carbonate. J Electrochem Soc 139:667–671
26. Doyon JD, Gilbert T, Davies G, Paetsch L (1987) NiO solubility in mixed alkali/alkaline earth carbonates. J Electrochem Soc 134:3035–3038
27. Kunz HR, Bregoli L J (1990) Ionic migration in molten carbonate fuel cells In: Selman JR, Shores DA, Maru HC, Uchida, I (eds) Carbonate fuel cell technology, (The Electrochemical Society PV90-16), pp 157–168
28. Veldhuis JB, Eckes FC, Plomp L (1992) The dissolution properties of $LiCoO_2$ in molten 62:38 mol% Li:K carbonates. J Electrochem Soc 139:L6–L8

29. Hatoh K, Niikura J, Yasumoto E, Gamo T (1994) The exchange current density of oxide cathodes in molten carbonates. J Electrochem Soc 141:1725–1730
30. Motohira N, Senso T, Yamauchi K, Kamiya N, Ota K (1999) Solubility of nickel in molten carbonates-The effect of Mg addition. In: The 6th FCDIC fuel cell symposium proceedings (Fuel Cell Development Information Center, Japan), pp 237–240
31. Yuh CY, Selman JR (1991) The polarization of molten carbonate fuel cell electrodes I. Analysis of steady-state polarization data. J Electrochem Soc 138:3642–3648
32. Morita H, Mugikura Y, Izaki Y, Watanabe T, Abe T (1997) Analysis of performance of molten carbonate fuel cell V. Formulation of anode reaction resistance. Denki Kagaku 65:740–746
33. Morita H, Mugikura Y, Izaki Y, Watanabe T, Abe T (1998) Model of cathode reaction resistance in molten carbonate fuel cells. J Electrochem Soc 145:1511–1517
34. Ramaswami K, Selman JR (1994) Rotating disk studies in molten carbonates III. Diffusion coefficients and bulk concentration in lithium carbonates. J Electrochem Soc 141:2338–2343
35. Vogel WM, Smith SW, Bregoli LJ (1983) Studies of the reduction of oxygen on gold in molten Li_2CO_3-K_2CO_3 at 650 °C. J Electrochem Soc 130:574–578
36. Malinowska B, Cassir M, Devynck J (1994) Design of a gold ultramicroelectrode for voltammetric studies at high temperature in glass-corrosive media (molten carbonate at 650 °C). J Electrochem Soc 141:2015–2017
37. Uchida I, Nishina T, Mugikura Y, Itaya K (1986) Gas electrode reactions in molten carbonate media I. Exchange current density of oxygen reduction in (Li + K)CO_3 eutectic at 650 °C. J Electroanal Chem 206:229–239
38. Lee CG, Nakano H, Nishina T, Uchida I, Kuroe S (1998) Characterization of a 100 cm^2 class molten carbonate fuel cell with current interruption. J Electrochem Soc 145:2747–2751
39. Yuh CY, Selman JR (1988) Characterization of fuel cell electrode processes by AC impedance. AlChE J 34:1949–1958
40. Morita H, Mugikura Y, Izaki Y, Watanabe T (1999) Analysis of performance of molten carbonate fuel cell VI. Analysis of Nernst Loss on current interrupt wave. Electrochemistry 67:438–444
41. Lee CG, Lim HC (2005) Experimental investigation of electrode reaction characteristics with reactant gas addition measurement in a molten carbonate fuel cell. J Electrochem Soc 152: A219–A228
42. Lee CG, Nakano H, Nishina T, Uchida I, Izaki Y, Kuroe S (1996) Transient response analysis on an 100 cm^2 class molten carbonate fuel cell. Denki Kagaku 64:486–490
43. Morita H, Nakano H, Mugikura Y, Izaki Y, Watanabe T, Uchida I (2003) EIS as a tool to determine fuel flow distribution in molten carbonate fuel cells. J Electrochem Soc 150:A1693–1698
44. Lee CG, Hwang JY, Oh M, Kim DH, Lim HC (2008) Overpotential analysis with various anode gas compositions in a molten carbonate fuel cell. J Power Sources 179:467–473
45. Lee CG, Yamada K, Hisamitsu Y, Ono Y, Uchida I (1999) Kinetics of oxygen reduction in molten carbonates under pressurized Air/CO_2 oxidant gas conditions. Electrochemistry 67:608–613
46. Tomczyk P (2006) MCFC versus other fuel cells—Characteristics, technologies and prospects. J Power Sources 160:858–862
47. Lee CG, Kim DH, Lim HC (2007) Electrode reaction characteristics under pressurized conditions in a molten carbonate fuel cell. J Electrochem Soc 154:B396–B404
48. Lide DR (2007) Handbook of chemistry and physics, 88th edn. CRC Press, Boca Raton, pp 6–143

Monsoon Systems, Modeling of

Chien Wang[1], William K. M. Lau[2]

[1]Department of Earth, Atmospheric, and Planetary Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA

[2]Laboratory for Atmospheres, NASA Goddard Space Flight Center, Greenbelt, MD, USA

Article Outline

Glossary

Aerosol Small particles suspended in the atmosphere in solid or liquid phase.

El Niño and Southern Oscillation (ENSO) Two intimately linked phenomena in tropical regions; El Niño (“the Christ Child” in Spanish) refers to the significant increase in sea surface temperature that irregularly occurs during Christmas time over eastern and central Pacific Ocean; Southern Oscillation refers to the low-latitude oscillation of sea level pressure centered respectively in the eastern Pacific and the western Pacific to Indian Ocean.

General circulation model (GCM) A computer program that solves numerically the time-dependent governing equations describing the evolution of atmospheric or oceanic circulation.

Intertropical convergence zone (ITCZ) A longitudinally extended zone near the equator that separates the northeast wind in the Northern Hemisphere from the southeast wind in the Southern Hemisphere near the Earth's surface.

Madden–Julian oscillation (MJO) An oscillation of zonal wind in both the boundary layer and upper troposphere propagating eastward with an average speed of 5 m/s across equatorial Indian and western and central Pacific Ocean.

Moist static energy (MSE) An atmospheric thermodynamic variable defined as:

$$\mathrm{MSE} = C_p T + gz + L_v q$$

Here C_p is the specific heat of air, T is air temperature, g is gravity, z is height above surface or a given reference level, L_v is the latent heat of water vaporization, and q is the ratio of water vapor to total air in mass.

Tropical biennial oscillation (TBO) A zonal wind oscillation in the equatorial stratosphere.

Definition of the Subject

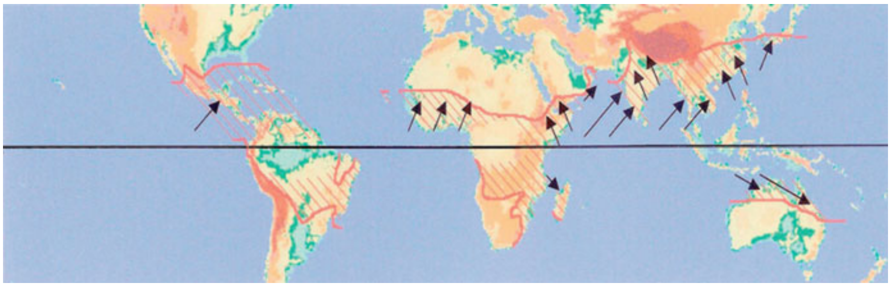
The word monsoon derives from the Arabic word "mausim," referring to the seasonal reversal of prevailing low-level winds blowing from relatively cold and moist ocean to warm land during the wet season (summer), and from cold and dry land to ocean during the dry season (winter). Monsoon systems are found in tropical regions from Africa, India, East Asia, Australia, and the Americas. Deep convection along with heavy rainfall occurs during the wet monsoon season over land as well as ocean. Typically, in monsoon regions the rainfall in wet season accounts for more than half of the annual surface precipitation. Monsoon evolution heavily influences human activities including agricultural practice and societal habits of billions of people living in monsoon regions. Knowledge and improving prediction of the onset, maintenance, variability, and key drivers are critical to the livelihood of these people. Because the monsoon is an integral part of the global climate system, better understanding of the monsoon is pivotal to predict future climate change and also the response of monsoon systems to such change.

The onset and strength of monsoons are determined by dynamical and thermodynamical processes not only locally over monsoon regions but also remotely over other regions. Computer models combining related dynamical, physical, and chemical processes in various scales are hence important tools to examine the current understanding of monsoon dynamics. These models can be used to test various hypotheses, and to actually simulate and forecast monsoon evolution. Simple models used in the earlier days of monsoon research mostly described the monsoon system from an energy budget perspective. These models could capture very rudimentary features of monsoon energy conversion. However, they lack the capability to go further in revealing the details of rainfall intensity and distribution, and particularly the timing of monsoon onset. Sophisticated three-dimensional regional and global climate models have been used in recent years to simulate monsoon systems, and to study the sensitivity of monsoon circulation and precipitation to various factors, including identification of the anthropogenic impact on monsoon system. These models have also been used to project monsoon evolution under different scenarios of possible future climate change.

Introduction

Over 60% of the world's population lives in monsoon regimes with a clear annual cycle of wind and precipitation. Such a cycle consists of a wet and a dry season. Wind in the lower atmosphere blows poleward from a relatively cold ocean to warm land during the wet season, and goes in the opposite direction during the dry season (Fig. 1). The monsoon regions include a large part of tropical and subtropical Asia and Africa as well as Australia, where some of the most populous nations in the world are located.

Agricultural activities, water resources, and many societal events in regions with a monsoon climate are strongly influenced by the wind reversal and the uneven distribution of rainfall. Forecasting monsoon rainfall and onset, the sudden transition from dry condition into a heavy downpour, has practical meanings to human activities in these places. In addition, whether future climate change would alter the behavior and strength of monsoon is a critical issue in making climate related strategies. The achievement of an adequate skill to forecast future monsoon evolution



Monsoon Systems, Modeling of. Figure 1
Land areas that have the majority of their rainfall in summer, associated with the poleward motion of deep convection. Where appropriate, low-level wind directions that carry moist warm air are indicated. In areas where there are no arrows, winds are relatively dry, or are weak (as over South America). *Shaded areas* show the normal maximum extent of deep convection (From [1] by J.F.P. Galvin with permissions from the author and Wiley)

relies on a good understanding of fundamental monsoon dynamics along with its variability. This requires knowledge about the formation mechanism and the major driving factors, including both natural and anthropogenic ones, of the monsoon system.

Monsoon Dynamics Fundamentals

Research to identify the driving forces of monsoon onset and strength has been conducted by correlating various diagnostic quantities with monsoon system characteristics. These characteristics include moisture, clouds, precipitation, and the large-scale circulation. With continued advancement of observational technology, from rain gauge stations to satellite monitoring, such effort has gained momentum, leading to improved knowledge. However, improved knowledge often reveals additional complexity of monsoon systems. This will in turn require even better understanding to further improve theory and modeling. This requirement has led to the building of a hierarchy of regional through global climate-system models for monsoon research.

It has long been held that differential heating on land and ocean following solar insolation cycle to be the major formation mechanism of the large-scale monsoon circulation [2, 3]. Such a heating contrast would force wind blow toward warm region, although because of the geostrophic constraint the actual wind direction is altered. Various rather simple models were developed to simulate monsoon systems based on the differential heating concept. These include zonally symmetric and other types of two-dimensional models that describe the zonal circulation and precipitation from ocean to land over monsoon regions, and often include a description of the planetary boundary layer. These simple models along with their limitations have been discussed extensively in literature [4, 5].

Attempts to forecast monsoons have linked monsoon strength with other phenomena or processes, ranging from the snowfall on the hills of Himalaya in the previous winter [6], mountains [7, 8], to the El Niño and Southern Oscillation (ENSO) ([9] and many others). Some of these factors are still within the framework of the large-scale land-sea thermal contrast model while others clearly connect to global climate dynamics.

In recent years, there have been studies suggesting that as a northward extension of the Intertropical convergence zone (ITCZ), the onset of the monsoon could just be a result of a longitudinal sea surface temperature (SST) gradient, not necessarily the traditionally held land-ocean temperature gradient (e.g., [10, 11]). It has also been demonstrated that the poleward boundary of monsoon circulations are co-located with a maximum in sub-cloud layer moist static energy (or entropy; MSE), corresponding to the minimum of vertical meridional wind shear [12–14]. Such a location and extent of the monsoon would also be influenced by the position of subtropical thermodynamic forcing as well as the advection of MSE.

Due to a meteorological phenomenon called the "thermal wind balance," the heating over land to the north and the cooler ocean to the south during the summer monsoon will produce easterly winds aloft and westerly winds below. When the upper troposphere easterly wind is strong, in the case of a strong monsoon, instability of easterly jet may stimulate the formation of eddies. The formation of these eddies might not always amplify the monsoonal circulation because the northeastward flow from ocean could bring low MSE air to land [5, 12, 13]. A recent proposal [15] actually suggested viewing monsoons as eddy-mediated transitions in the tropical overturning circulation between regimes that are distinct in the degree to which eddy momentum fluxes control the strength of the circulation. In this study, the idealized general circulation model (GCM) simulation on an aqua-planet demonstrated that the role of land in monsoon onset is to just provide a media of low thermal inertia. Whenever such surface differences in heat capacity exist monsoon onset would happen regardless of other surface inhomogeneities. Therefore, interactions between extratropical eddies and the tropical meridional overturning circulation could be essential for monsoons. In addressing the interaction of monsoon and other dynamical system, it was indicated that the feedback of atmosphere to SST forcing might have played a critical role in monsoon evolution [16].

One specific implication of these new hypotheses is on the predictability of the monsoon system. It has been argued that because the dominant forcing of monsoon system are the rather slow processes that control the tropical sea surface temperatures (SST), therefore, the predictability of monsoon rainfall at least in monthly or seasonal scale may be promising [17]. However, should the extratropical eddies and atmosphere to ocean feedback be critical in monsoon onset and evolution, the monsoon predictability issue even on relatively long time scales would be much more complicated. The predictability of the monsoon is further confounded by the ubiquitous presence of monsoon-intraseasonal oscillations (MISO) in both the summer and the winter seasons. These are intrinsic oscillations in the monsoon region, with characteristic timescales of 20–70 days, arising from the organization of tropical convection over the ocean associated, (e.g., planetary scale Madden–Julian Oscillation (MJO); see [18] and many others). MISO are mediated by SST changes as well as the monsoon regional topography, and influence the onset, break, and maintenance of the monsoon as well through interactions with ENSO. Realistic simulations of MISO and MJO have been a challenge even for the state-of-the-art climate models.

Modeling the Monsoon

Because of the complex, multi-scale characteristics of monsoon systems, efforts to understand and to examine various hypotheses about monsoon onset, evolution, and strength have to rely largely on computer models in combination with available data. In order to understand the interaction between this large-scale moist circulation and many other complicated but critical processes, ranging from ocean–atmosphere interaction, extratropical–tropical interaction, MISO, to potential “teleconnection” through synoptic waves between tropical systems in distance, one needs to use a three-dimensional global climate model or a regional climate model interacting with a global climate model.

Simulation of monsoon systems using three-dimensional global models started from the very early stage of atmospheric general circulation models. Typical model used in these early simulations had coarse horizontal resolution (270–540 km) and 11 vertical layers, forced by prescribed seasonal variations of insolation and often sea surface temperature [19]. These simulations were mostly used for exploratory purposes due to their short integration time (often shorter than 3 years), coarse model resolution, and the prescription of some fields of potential importance in modeling the monsoon. Understanding the onset of the monsoon was clearly a far-reach at that stage.

The availability of multi-decadal sea surface temperature data allowed three-dimensional atmospheric general circulation models (AGCM) to simulate monsoon system evolution driven by observed SST time series. This type of simulations follows the procedure of the Atmospheric Modeling Intercomparison Project (AMIP, and AMIP II later; [20]), forced by “real-time” SST data and thus ignored the feedback between the atmosphere and ocean. AMIP models, which generally include interactive land surface models, have the advantage of identifying atmospheric feedback mechanisms without dealing with the complexity of the coupled ocean–atmosphere system. They were the models of choice in the 1980s–1990s. Nowadays, long-term climate simulations are generally done with coupled ocean–atmosphere models. However, AMIP models are still useful when run at high resolutions to test sensitivity to model atmospheric processes of physics, chemistry, and aerosols, and interaction of the atmosphere with surface vegetation. Some current AMIP runs are conducted at mesoscale resolution (<25 km) globally, and others are configured even at higher resolution for global hurricanes studies. With such configurations, models could explore the onset of the monsoon and aspects of the MJO and MISO [21]. The very high-resolution AMIP models are extremely computationally demanding, and can only be run using high-performance computers at large institutions.

Studies exploring multi-decadal to centennial timescale issues that need to consider the role of atmosphere-ocean feedback in monsoon dynamics, typically utilize moderate-to-low resolution (100–200 km) AGCMs coupled with either mixed-layer ocean model or full ocean general circulation model to reduce the computational demand. Regional climate models have also been used for this purpose. The influence of future climate change on monsoon evolution has also been studied mostly using the ensemble results of three-dimensional climate model simulations included in the Fourth Assessment Report (AR4) of the Intergovernmental Panel of Climate Change (IPCC). A fast growing effort in recent years is to study the impacts of anthropogenic forcings particularly of aerosols on monsoon circulation and precipitation.

This entry will begin by describing efforts to use climate models to simulate monsoon systems. Recent research on the potential role of aerosols in monsoon systems is then described. The projections of monsoon system changes under possible climate change scenarios will also be discussed, concluding with an overview of the future opportunities. The discussion will be focused on the utilization of general circulation models and regional climate models, of moderate-to-low resolution in simulating monsoon systems and the study of the sensitivity of monsoon to various climate dynamical processes as well as anthropogenic impacts on equilibrium climate, or on climate time scales of a century or less.

Simulating Monsoon Systems Using Climate Models

Before attempting to use a computer model to forecast monsoon evolution, one would ask the very question that how well the model might reproduce the major observed characteristics and variability of a monsoon, if some of the known factors controlling monsoons were included in the model (of course this condition itself is somewhat a unsettled issue). This actually leads to a type of modeling study, so-called retrospective modeling. In modeling monsoon systems, a retrospect simulation would be performed by prescribing the time series of sea surface temperature, assuming that the ocean is such a large heat reservoir comparing to the atmosphere so that the change in SST reflect the long-term state of energy balance. This type of modeling allows modelers to concentrate on issues other than the feedbacks from the atmosphere to ocean. With the assumption that historical SST change might well represent the effect of all the long-term forcings, this type of simulations is expected to capture major features of the monsoon systems in the past.

In simulating monsoon systems, it is essential for the model to capture certain representative features of the system. These would at least include the reverse atmospheric circulation between the upper and lower levels, the onset of monsoon rainfall, and total precipitation during monsoon season. In addition, the climatological rainfall patterns including land-ocean partition during monsoon is also among important system characters. A more subtle and difficult task in modeling is to capture the weak correlation between ENSO and Indian summer monsoon rainfall [9], and the correlation between anomaly of Gulf of Guinea SST and the dipole rainfall pattern of the West African monsoon over Guinean coast and the Sahel [22]. Interannual and decadal variability is another important test for both retrospect modeling and for revealing the dependency of monsoon systems on critical forcings. Simulating the intraseasonal variation of detailed rainfall strength and distribution (e.g., [23]) would be an important task for regional or high-resolution global models.

Much progress has been made through years of efforts in modeling the monsoon system. In early stage of such attempt, models typically had low resolution and prescribed seasonal forcing of SST. Arguably, the physics processes such as clouds and radiation in those models were also poorly treated comparing to models used today. Nevertheless, the early models captured certain basic features of the monsoon system (mostly on Indian summer monsoon due to its rather clearly defined annual cycle and relatively extensive analyses) such as the reverse low-level circulation over northern Indian Ocean (e.g., [19]). Besides, the role of certain hypothesized driving factors of monsoon circulation such as mountains [7] and the anomaly of Arabian Sea surface temperature [24] had been also examined. The simulated onset of monsoon, however, was much delayed and the distribution of rainfall and intensity has large biases compared to observations.

With the availability of decadal-long observed SST dataset, nearly all the major AGCMs in the world joined the effort of Atmospheric Modeling Intercomparison Project (AMIP) in the 1990s. An AMIP configuration is a typical retrospect simulation, where atmospheric general circulation models were driven by a time series of observed SST data to reproduce the past climate. Lau and Yang had used the Goddard Space Flight Center GCM (4 × 5 degree resolution along latitude and longitude, respectively and 17 vertical layers) in an AMIP 1979–1988 simulation to examine the Asian monsoon system [25]. The model was able to capture many broad-scale structures of Asian monsoon system, including evolution of global and regional circulation, rainfall, moisture flux, and intraseasonal and synoptic variability. Interestingly, the model also successfully simulated multiple onset of East Asian monsoon along with the onset of Indian summer monsoon. These onsets were initiated by a sudden jump of the ITCZ from the equator to 10°N, related to a northward shift of the ascending branch of the local Hadley circulation. Clearly, this was a significant advance from the early simulations. On the other hand, the model did not reproduce observed rainfall distribution and quantity over many precipitation centers. Intraseasonal transition of ITCZ between equatorial region (ocean) and monsoon land was not well captured. The East Asian monsoon trough was also severely underdeveloped in the model. These shortcomings actually existed in most AMIP models.

In AMIP-type simulations, the atmosphere-ocean feedback is set aside. The modeling focus is instead on the atmospheric simulation, presuming that the SST time series realistically reflects the forcing of the past. Based on the results of ensemble AMIP simulations, Wang et al. indicated that a lack of the atmospheric feedback in simulations forced by observed SST could lead to serious biases in modeled monsoon precipitation [16]. They found that the atmospheric feedback to SST forcing is more significant than SST to atmospheric forcing. Therefore, coupled model would be critical in even retrospect modeling of monsoon and the atmospheric feedback to tropical SST forcing needs to be included. Meehl et al. further demonstrated an improvement of modeled monsoon features using a higher resolution (T85) coupled atmosphere-ocean model compared to a lower resolution (T42) AMIP configuration model [26]. There have also been reports of significant improvement in modeling monsoon features made simply by using high-resolution atmospheric GCM or regional climate model driven by observed SST (see [27]).

When using coupled model to simulate monsoon system, drifts of SSTs away from observation could become an issue. With a rather coarse-resolution coupled model (4.5 × 7.5 degree and nine layer for the atmospheric model; 5 × 5 degree and four layer for the ocean model), Meehl indicated that without adopting correction terms to force the coupled model to the observed state, model-simulated SSTs in the tropics tend to be too cold. This bias would enhance land-sea temperature contrast in the monsoon region, yet the pattern of mean monsoon seasonal precipitation and the variability of the simulated South Asian monsoon (SAM) were comparable to the observed pattern [28]. One alternative method to the AMIP configuration is to predict SST using a 2½-layer tropical ocean model between 30°S and 30°N and to prescribe SST in other places [16]. The SST-monsoon rainfall correlations indicated by observations (with 1 month lag) were reflected correctly in the simulation conducted by using this method.

Models have also been used to identify certain hypothesized driving factors behind monsoon variability. For instance, it is known that the Tropical Biennial Oscillation (TBO), a variation in precipitation occurring with approximately a 2-year period, affects monsoon strength. Therefore, identifying the relative importance of various potential conditions leading to TBO transitions could help us to understand the factors that affect monsoon variability. It was found that among three conditions hypothesized to contribute to TBO transitions, tropical Indian Ocean SST anomaly and tropical Pacific Ocean SST anomaly are more effective than anomalous meridional temperature gradients over Asia [29]. The two types of tropical SST anomalies were found to dominate the TBO transitions and thus produce large monsoon response in the model sensitivity results. In addition, the location of the SST anomalies over the tropical Indian Ocean is found to be important. Warm SST anomalies throughout the tropical Indian Ocean enhance rainfall over the ocean and South Asian land areas. Warm SST anomalies near equatorial Indian Ocean produce increased rainfall locally with decreased rainfall over South Asian land areas.

Despite significant progresses achieved, there is still much room for improvement regarding the performance of current climate models in simulating various features of monsoon systems from mean state to variability [30]. For instance, among 18 coupled GCMs that participated the effort of the IPCC Fourth Assessment Report (IPCC AR4), only six of them were found to have realistic representation of South Asian monsoon precipitation climatology in the twentieth century [31]. It is noteworthy that these six models all had large pattern correlation and small root-mean-square differences (RMSD) with observations in modeling June–July–August–September (JJAS) rainfall climatology, both over India (7°–30°N, 65°–95°E) and for the larger monsoon domain (25°S–40°N, 40°E–180°). Only four out of these six, though, exhibited a robust ENSO-SAM teleconnection.

Recent results from the West African Monsoon Model Evaluation (WAMME) project showed that AMIP-type models (both regional and global) generally have reasonable skills in simulating the pattern of the spatial distribution of West African monsoon (WAM) in seasonal mean precipitation, surface temperature, averaged zonal wind in latitude-height cross-section, and low-level circulation [32]. However, there are large differences among models in addition to model biases compared to observations in simulating spatial correlation, intensity, and variability of precipitation.

In a well-designed analysis [22], the abilities of above-mentioned 18 models were evaluated based on whether they can correctly simulate the circulation characteristics that support the precipitation climatology and the physical processes of a prominent mode of WAM variability, that is, the “rainfall dipole” variability that is often associated with dry conditions in the Sahel when SSTs in the Gulf of Guinea are anomalously warm. It was found that each model captured the largest-scale rainfall pattern featuring a zonally oriented precipitation maximum, but about one third of them did not generate the West African monsoon, that is, they did not bring the ITCZ and its associated rainfall onto the African continent during boreal summer. Only three further captured the three precipitation maxima over the continent, that is, the maximum on the west coast, over the eastern portion of the Guinean coast, and over the Ethiopian highlands. It was thus concluded that the current generation of coupled GCMs is much more capable of accurately representing the summer precipitation climatology over North America and Europe than over Africa.

In modeling the Sahel drought during 1970s–1980s, the most pronounced climate signal in WAM regions that had been suggested as a consequence of warm anomalous SST surrounding Africa (e.g., [33]), Lau et al. evaluated the performance of 19 coupled general circulation models (also AR4 models) in twentieth-century simulations [34]. They found that only eight of these models produced a reasonable Sahel drought signal, while others either produced excessive rainfall over the Sahel during the observed drought period or showed no significant deviation from normal. Even the model with the highest prediction skill of the Sahel drought could only predict the increasing trend of severe drought events but not the beginning and duration of the events. Based on the analysis, it was recommended that in order to accurately simulate the Sahel drought, models need to have a strong coupling between Sahel rainfall and the SSTs of both Indian and Atlantic Ocean, in addition to a robust land surface feedback with strong sensitivity of precipitation and land evaporation to soil moisture.

The performance of 12 coupled models in the Coupled Model Intercomparison Project phase 3 (CMIP3; the same group of models that participated in IPCC AR4) in simulating present-day East Asian monsoon has been examined as well [35]. Almost all of these models were found to be able to reproduce observed interannual variability of summer rain belt and associated circulation. The models can also reproduce the interannual variation of the western North Pacific subtropical high (WNPSH) in the lower troposphere, a parameter closely related to the interannual variation of summer rainfall. However, the predicted quantities of interannual variation of WNPSH from these models differ significantly.

Modeling the Impacts of Aerosols on Monsoon System

Atmospheric aerosols serve as a critical player in the climate system. All aerosols attenuate solar radiation through either scattering or absorption, both leading to cooling at the Earth's surface. In addition, absorbing aerosols warm the atmosphere, affecting atmospheric profile and thus dynamical processes. Aerosols also dominate the cloud formation in the atmosphere, serving as cloud condensation nuclei (CCN) or ice nuclei (IN) to provide preexisting surfaces and thus a superior mean for cloud particles to form than homogeneous nucleation. Therefore, changes in aerosol properties such as number concentration, size distribution, or chemical composition (hygroscopicity) are expected to affect atmospheric systems from regional to global scales including the monsoon.

Human activities produce aerosols containing inorganic matters such as sulfate and nitrate, and organic matter as well as black carbon. These anthropogenic aerosols are regarded as an addition to the natural aerosols that mainly include dust, biogenic, and sea salt particles, and hence exert a forcing to the climate system. Studies suggested that the reduction in Indian monsoon strength in recent decades could be a result of an increase of anthropogenic aerosols over monsoon regions, mostly coinciding with the fastest growing economies including China and India as well as Southeast Asia [36].

On the other hand, a recent analysis of 1951–2003 daily gridded rainfall data over India revealed a decreasing trend in both early and late monsoon rainfall and number of rainy days, implying a shorter monsoon over India [37]. There is also a sharp decrease in the area that receives a certain amount of rainfall and number of rainy days during the season. An increase in the frequency of heavy precipitation in the Indian summer monsoon was also identified [38].

A great deal of attention has been paid to the influence of anthropogenic aerosols (particularly absorbing aerosols) on tropical precipitation in recent years. Absorbing aerosols influence the climate in distinctly different ways from aerosols that primarily scatter energy back to space. Studies using different general circulation models all indicate that direct radiative forcing (DRF) of absorbing black carbon (BC) aerosols can lead to a northward shift of precipitation in ITCZ over the Pacific Ocean [39–41]. Modeling studies also suggest that DRF of aerosols could have a significant impact on the monsoon systems as well [42]. Correlations between estimated precipitation/circulation changes with increasing trend of aerosols have unquestionably fueled the researches toward this direction.

Studies of aerosol-monsoon impact are rapidly growing not only for the Indian summer monsoon, but also for the East Asian monsoon, the West African monsoon, and the Australian monsoon. Most of these studies are conducted by using three-dimensional atmospheric GCMs or coupled climate models. Paired simulations driven respectively by including and excluding aerosol effects, or by including a reference and an altered aerosol profile along with aerosol effects, provide a comparison in climate response between different aerosol forcing assumptions. The aerosol effects would be isolated barring the assumption that model's artifact in simulating the monsoon system would not be significantly amplified by using different aerosol profiles. The descriptions of aerosol and aerosol–climate interaction vary in studies though.

Impacts of Aerosols on the Asian Monsoon

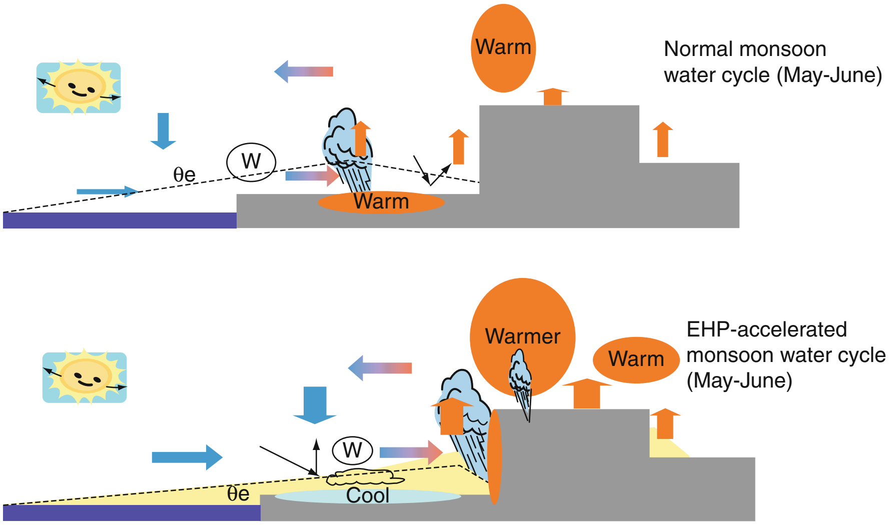
Earlier works focused on impacts of absorbing aerosols (black carbon and dust) on atmospheric water cycle in the Asian monsoon, by using prescribed aerosol distributions from global chemistry transport models and/or observations. These studies excluded the dynamical feedbacks between winds and precipitation features and the aerosol distribution. An early exploratory study tested the climate response to absorbing aerosols over China and India [43]. In this study, a 12-layer and

4 × 5 degree resolution three-dimensional AGCM was used to explore the model response to prescribed aerosol optical depth and single scattering albedo over China and India only. The researchers found that the convection would be enhanced along 20°–30°E in longitude from eastern China to Indian subcontinent in responding to the added aerosol forcing. Despite many discrepancies in detailed results between this study and later ones, perhaps attributed to the regional-only aerosol loading and rather coarse resolution of the model in [43], the general response in large-scale dynamics associated with the monsoon systems caused by the direct radiative effects of absorbing aerosols remains consistent with later studies. For instance, Wang noticed an enhancement of the Indian summer monsoon circulation by the direct radiative forcing of black carbon aerosols in a coupled GCM simulation, though the analysis was done on an annual-mean base so that the seasonal features of the circulation were not discussed [39, 44]. A similar effect of BC aerosols was also found in another study, though where the simulation was driven by prescribed SST [45].

Perhaps the most interesting outcome in recent modeling efforts of aerosol-monsoon studies is the proposals of various hypotheses on the mechanisms of aerosol impact specifically on Indian summer monsoon. The discussions are also centered at the role of absorbing aerosols.

The radiative effects of absorbing (primarily dust and anthropogenic carbonaceous) aerosols in cooling the surface (dimming effect) and in heating the atmosphere can play different roles in affecting the monsoon system. The cooling over land from absorbing aerosols would assist lowering the land-ocean temperature gradient. Ramanathan et al. found that an increase in the BC DRF over Indian Subcontinent and surrounding regions in their model leads to a reduction of monsoon precipitation while an enhancement to the pre-monsoon precipitation of March–April–May (MAM) [36]. Using a coupled atmosphere-ocean general circulation model with prescribed black carbon direct radiative forcing, Meehl et al. found similar circulation and precipitation changes due to BC impact in the pre-monsoon (enhancement) and in monsoon season (reduction) [46]. It was also found that although during the monsoon months the effect of BC is likely to reduce the precipitation over India, it might enhance the precipitation over the elevated Tibetan Plateau. Meehl et al. suggested that BC DRF could weaken the surface temperature gradient between the tropical waters and the land of the Indian Subcontinent. This could serve as the forcing mechanism of BC on the monsoon circulation and precipitation, that is, through the dimming effect.

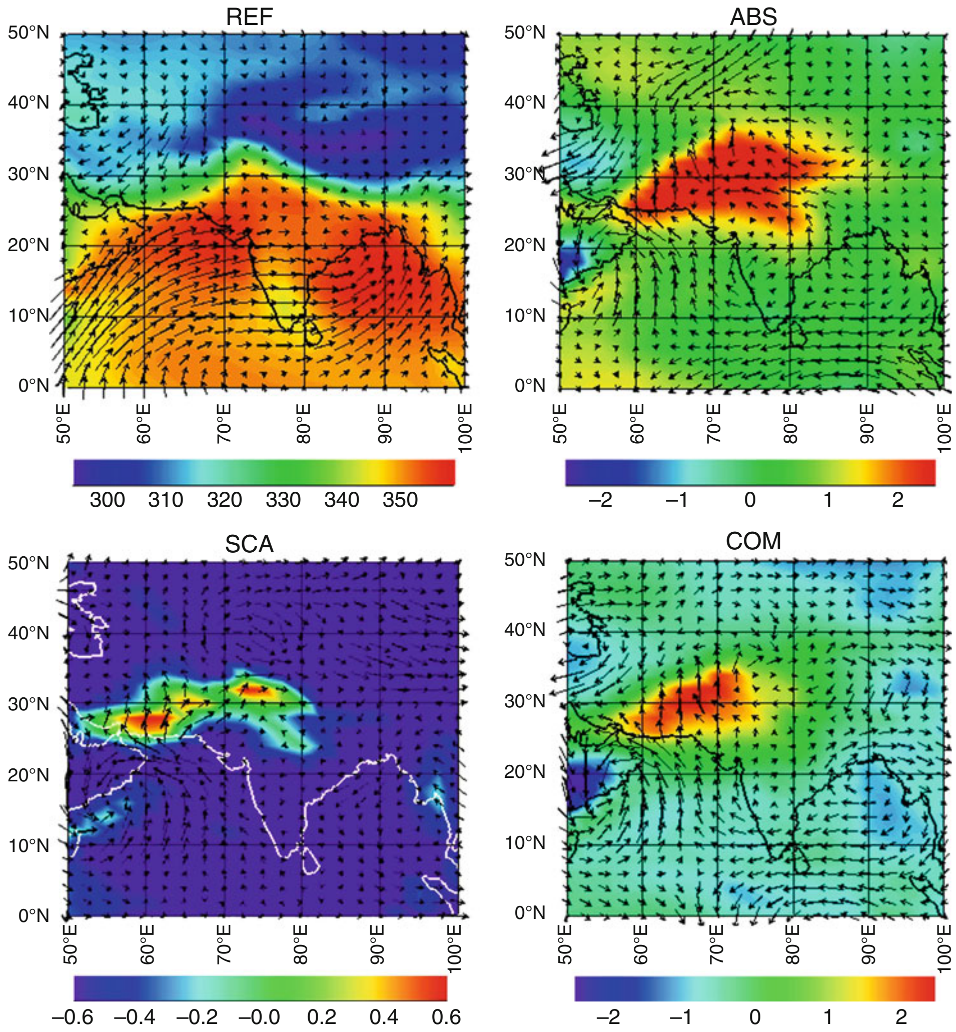
One specific characteristics of absorbing aerosols is its heating to the atmosphere. How would this effect play a role in aerosol-monsoon impact is also discussed. Lau et al. used an atmospheric GCM driven by prescribed global three-dimensional climatology of aerosol optical depth to examine the direct effects of aerosol on the monsoon water cycle variability [34]. The study suggested that, referred to as an “elevated heat pump” effect (EHP) (Fig. 2), dust mixed with black carbon aerosols that extend against the foothills of the Himalayas over the Indo-Gangetic Plain (IGP) in April and May could heat the air. This would initiate a positive feedback by drawing water convergence from oceans first and then form condensation and thus further heating over the slope of the Plateau. Based on this hypothesis, monsoon precipitation would be suppressed over central India due to aerosol-induced surface cooling. However, precipitation would come earlier and be enhanced over northern India and the southern slope of the Tibetan Plateau. Monsoon rainfall in July and August over the entire India would also be enhanced. Lately, using satellite aerosol index (AI) data and observed clouds and precipitation data, two studies have demonstrated the existence of anomalous absorbing aerosol loading in late spring over IGP [47, 48]. Both works also suggested a correlation of this aerosol anomaly with variation of monsoon evolution. A widespread warming over the Himalayan-Gangetic region and consequent strengthening of the land-sea thermal gradient was also found recently through satellite microwave sounding data [49]. This trend is most pronounced in the pre-monsoon season, resulting in a warming of 2.7°C in the record. All these observation-based analyses appear to be consistent with the EHP effect. The hypothesis involves both atmospheric-heating and surface-cooling effect of absorbing aerosols as well as induced changes in cloudiness. It is different than the hypothesis that only emphasizes surface cooling. The feedback mechanisms introduced by EHP hypothesis are, however, more complicated.



Monsoon Systems, Modeling of. Figure 2
Schematic showing the monsoon water cycle (*top*) with no aerosol forcing and (*bottom*) with aerosol-induced elevated heat pump effect. Low-level monsoon westerlies are denoted by W. The *dashed line* indicates magnitude of the low-level equivalent potential temperature θe. Deep convection is indicated over regions of maximum θe. (See text for further discussions) (Adopted from [42], © American Meteorological Society. Reprinted with permission)

For example, Prive and Plumb indicated that the equatorial oceanic air would have low moist static energy so that the low-level convergence might lead to a negative feedback to monsoon circulation [12, 13].

More recent works include the use of interactive aerosols in the models, that is, the dynamical feedback to aerosol distribution and forcing. For example, the model used in [50] includes a size- and mixing state-dependent aerosol module that is fully coupled with the climate model. Certain sophisticated aerosol microphysical and chemical processes including aging and coating of carbonaceous aerosols with sulfate, along with optical properties of these mixed aerosols are also included. Using this interactive aerosol-climate model coupled with a mixed-layer ocean model, Wang et al. proposed another possible mechanism that absorbing aerosols could affect on Indian summer monsoon [50]. The researchers find that absorbing anthropogenic aerosols, whether coexisting with scattering aerosols or not, can significantly affect the Indian summer monsoon system. This is drawn from a comparison of the results of three simulations. The first two simulations each only included absorbing aerosols (ABS) and scattering aerosols (SCA), respectively; the third one included both types of aerosols (COM). Aerosol-induced climate responses in each of these runs were derived by comparing results to a reference run that excluded the aerosol effect (REF). The similarity in aerosol-induced response between absorbing-aerosol-only case (ABS) and the case with both types of aerosols (COM) was identified. Results of both cases also differ sharply from that of the scattering-aerosol-only case (SCA). The researchers further identified that the influence of absorbing aerosols is reflected in a perturbation to the moist static energy in the sub-cloud layer, initiated as a heating by absorbing aerosols in the planetary boundary layer (Fig. 3). The perturbation appears mostly over land, extending from just north of the Arabian Sea to northern India along the southern slope of the Tibetan Plateau. As a result, during the summer monsoon season, modeled convective precipitation experiences a clear northward shift, coincidently in general agreement with observed monsoon precipitation changes in recent decades particularly during the onset season (Fig. 4). According to previous works, the northward extent of monsoon convection should collocate with the maximum sub-cloud layer MSE [12–14]. Therefore, a small perturbation in such a location could lead to an observable change in distribution of convection and heavy

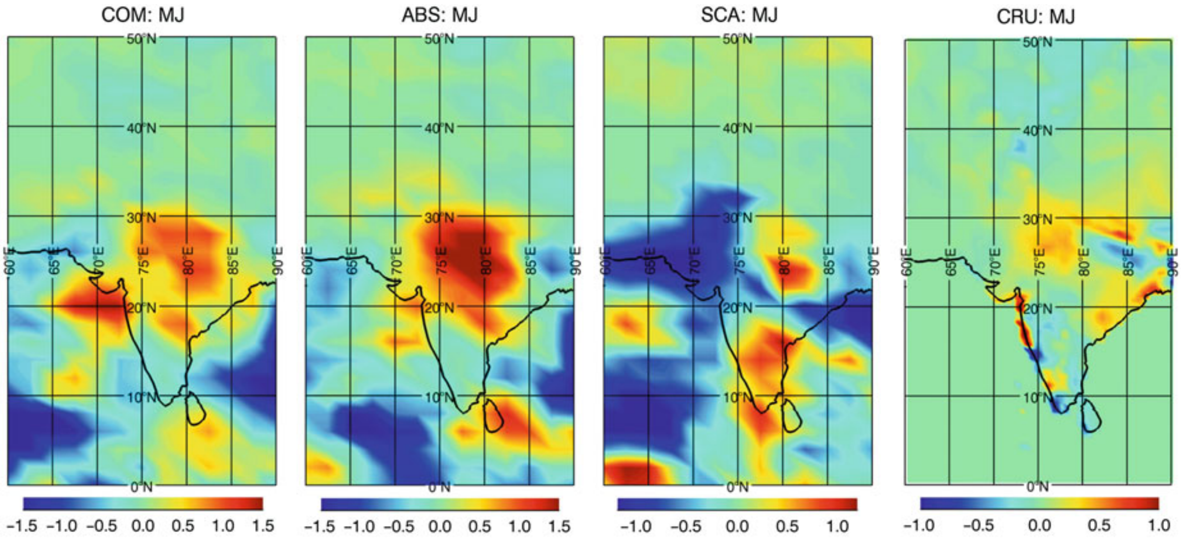


Monsoon Systems, Modeling of. Figure 3

May–June (MJ) mean of wind and moist static energy in the reference run (REF), which excludes the aerosol radiative effect, and anomalies of MJ mean wind and moist static energy derived from three model runs (ABS, SCA, and COM). Data shown are averaged values for the lowermost three atmospheric layers based on year 41–60 means. Unit wind vector = 1 m/s. Moist static energy is in 10^3 J/kg. Note that in the three anomaly plots, a different color scale is used in SCA. A terrain correction has been applied to the REF result (From [50], Copyright 2009 American Geophysical Union. Reproduced/modified by permission of American Geophysical Union)

precipitation. Interestingly, the modeled largest perturbation of absorbing aerosols on sub-cloud layer MSE appeared across this zone. Compared to the forcing required to significantly lower the meridional temperature gradient, the forcing of absorbing aerosols through perturbing MSE to influence monsoon dynamics and precipitation distribution is much more effective. The importance of sub-cloud layer processes, however, does not necessarily preclude the EHP that emphasizes the heating above the boundary layer. It is likely that the heating of the entire atmospheric column from the boundary layer to the upper troposphere could be important in creating the northward shift of the monsoon rainbelt.

At present, there is still a range of opinions about the reasons behind the impact of aerosols on Indian



Monsoon Systems, Modeling of. Figure 4
May–June average changes in convective precipitation (dm/season) derived from: COM, ABS, and SCA run for India and surrounding regions, and the observed precipitation change (land-only; dm/season) derived from the data of the Climate Research Unit (CRU) at the University of East Anglia. Model results shown are based on year 41–60 mean differences with REF run. CRU results are derived from differences between 20-year means of 1981–2000 and 1946–1965, and based on the version 2.1 dataset with 0.5° (From [50], Copyright 2009 American Geophysical Union. Reproduced/modified by permission of American Geophysical Union)

summer monsoon [42]. Detailed mechanisms of the aforementioned hypothetical impacts still remain to be examined.

Impacts of Aerosols on the West African Monsoon

Aerosol impacts on another monsoon system, the West African Monsoon (WAM) have also been studied. It is known that aerosols over this region are among the most abundant and persistent on the Earth with distinct seasonal variability. The dominant aerosol type is mineral dust from North Africa through May to August and biomass-burning smoke from southern Africa from July to September. The mixture of dust and biomass-burning smoke appear in November–February due to persistent yearlong dust emission from some North African sources and biomass burning in Sahel region [51–56]. Therefore, this region provides an ideal natural test bed for studying aerosol effects on precipitation. Analyses using satellite data have demonstrated that a high concentration of aerosols can induce a significant precipitation reduction in the WAM region along the coast of the Gulf of Guinea, particularly in the boreal late autumn and winter [57, 58]. A recent study [59] further compared the observational results to a global model simulation including only direct radiative forcing of black carbon [39]. It was found from both observations and model simulations that in boreal cold seasons anomalously high African aerosols are associated with significant reductions in cloud amount, cloud top height, and surface precipitation. This result suggests that the observed precipitation reduction in the WAM region is caused by radiative effect of absorbing BC. The mechanism for this reduction, however, remains to be revealed.

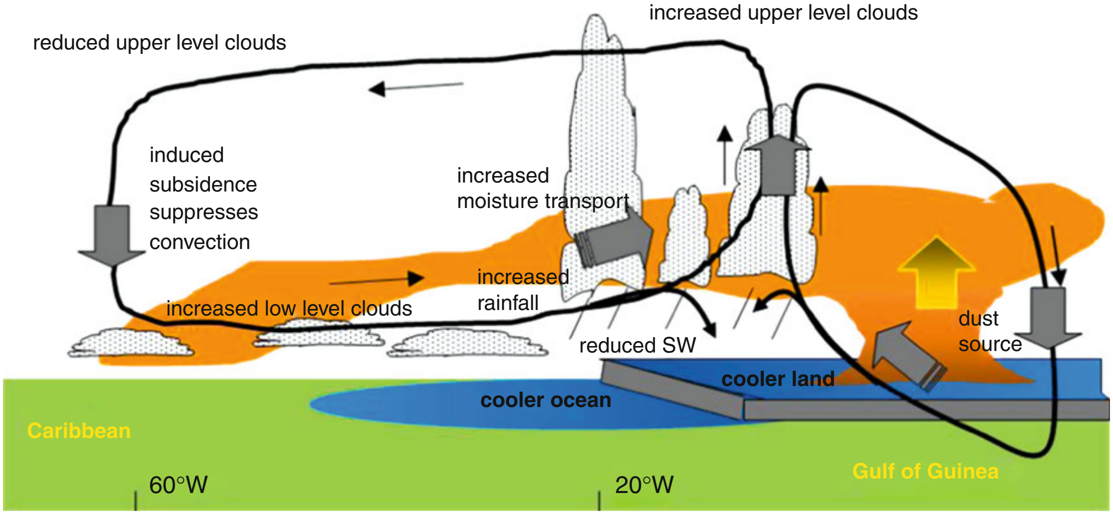
In connection to the hypothesis of aerosol-Indian summer monsoon effect proposed by Wang et al. [50], Eltahir and Gong found a correlation of the strength of the West African monsoon to subtropical meridional gradient of sub-cloud MSE [60]. Therefore, similar mechanism could also exist in aerosol-WAM effect.

Recently, Lau et al. showed from GCM experiment that the EHP effect by Saharan dusts and biomass-burning black carbon has a significant impact on the climate and water cycle of the North Atlantic and WAM [61]. They found that during the boreal summer, as a result of large-scale atmospheric feedback triggered by absorbing aerosols, rainfall and cloudiness are enhanced over the West Africa/Eastern Atlantic ITCZ while suppressed over the West Atlantic and Caribbean region. As shown in Fig. 5, the elevated dust layer warms the air over West Africa and the eastern Atlantic. As the warm air rises, it spawns an anomalous large-scale onshore flow carrying the moist air from the eastern Atlantic and the Gulf of Guinea. The onshore flow in turn enhances the deep convection over West Africa land, and the eastern Atlantic. The condensation heating associated with the ensuing deep convection drives and maintains an anomalous large-scale east–west overturning circulation, with rising motion over West Africa/eastern Atlantic and sinking motion over the Caribbean region. The response reflects a strengthening of the West African monsoon, manifested in a northward shift of the West Africa precipitation over land, increased low-level westerly flow over West Africa at the southern edge of the dust layer, and a near surface westerly jet underneath the dust layer over the Sahara. The dust radiative forcing also leads to significant changes in surface energy fluxes, resulting in cooling of the West African land and the eastern Atlantic, and warming in the West Atlantic and Caribbean. The EHP effect is most effective for moderate to highly absorbing dusts, and becomes minimized for reflecting dust with single scattering albedo at 0.95 or higher.

Additionally, from the same experiments the authors found strong modulation of the diurnal cycle and a northward shift of the African easterly jet, in conjunction with increased cyclonic vorticity to the south of its axis, and increased rainfall in the Sahel [62]. These modeling results are consistent with recent observations [63] showing that during the periods of a strong Sahara dust outbreak, the Atlantic ITCZ tends to be shifted northward of its climatological position, accompanied by a similar shift of the Africa easterly jet.

Impacts of Aerosols on the Australian Monsoon

The anthropogenic aerosol level in the Southern Hemisphere is lower compared to the condition of Northern



Monsoon Systems, Modeling of. Figure 5
Schematic diagram showing key features in the latitude domain 5–15°N, associated with the "elevated heat pump" mechanism by radiative heating of Saharan dust: anomalous Walker-type and Hadley-type circulations, increased moisture transport from the eastern Atlantic and the Gulf of Guinea to West Africa; enhanced rainfall over the Sahel and the ITCZ off the cost of West Africa; subsidence and suppressed cloudiness in the central Atlantic and Caribbean, and the Gulf of Guinea; cooling of the WAM land, and the upper ocean in the eastern Pacific underneath the dust plume (Adopted from [61] © Author(s) 2009, distributed under the Creative Commons Attribution 3.0 License)

Hemisphere. Therefore, the impact of local aerosols on Australian monsoon system is expected to be insignificant. However, since monsoon systems are closely associated with large-scale circulation, aerosol effects in the Northern Hemisphere could influence southern hemispheric circulation and thus precipitation by altering the general circulation patterns. This has been suggested in a recent modeling study [64]. Drawn from the results of a pair of ensemble simulations conducted using a coupled atmosphere-ocean climate model respectively including and excluding Asian aerosols, the researchers hypothesized that Asian aerosols could lead to an increase in both rainfall and cloudiness particularly over northwest Australia, which coincides with observed rainfall trend in the region since 1950s. The study suggested that this effect could be implemented through an altered latitudinal gradient of temperature (and thus of pressure) over tropical Indian Ocean by Asian aerosols, which would further enhance monsoonal circulation toward Australia. A recent analysis of the twentieth-century modeling results of 24 CMIP3 models, all including either only direct or both direct and indirect aerosol forcing, however, cannot provide support to the above hypothesis [65]. Despite of the inclusion of aerosol effect in these models, their ensembles did not produce the hypothesized rainfall increase in northwest Australia.

Climate Change and Monsoon System

Analyses using oxygen isotope data from Chinese caves providing information about monsoons over millennia suggest that Asian monsoons are influenced by changes in summer insolation in the Northern Hemisphere [66, 67]. However, variability in shorter terms can also be influenced by other factors. Historically, such variation in Asian monsoon system might have triggered social unrest and thus played a key role in causing demise of several Chinese dynasties [68]. Similarly, the persistent drought in Sahel occurred later last century also led to serious food supply problems and could well be responsible for certain conflicts in Africa. It is thus critical to understand these variabilities of monsoon systems and to identify the natural and anthropogenic influences on such variations.

Several persistent trends have been revealed recently. The reconstructed monsoon winds for the past 1,000 years using fossil *Globigerina bulloides* abundance in box cores from the Arabian Sea suggested an increase in strength during the past four centuries while the Northern Hemisphere has been warming [69]. This implies that the Indian summer monsoon strength could be enhanced during the coming century as greenhouse gas concentrations continue to rise and northern latitudes continue to warm. In the most recent decade, the sea surface winds over the western Arabian Sea have been continually strengthening [70]. Such escalation of summer monsoon winds, accompanied by enhanced upwelling, leads to an increase of more than 350% in average summertime phytoplankton biomass along the coast and over 300% offshore, implying that the current warming trend of the Eurasian landmass is making the Arabian Sea more productive.

Over India, analyses based on rainfall data since early 1950 suggest that in the last half century, the frequencies of moderate and low rain days over the entire country have significantly decreased while the frequency and the magnitude of extreme rain events has significantly increased [38, 71]. Decreasing trends were also found in both early and late monsoon rainfall and number of rainy days, implying a shorter monsoon over India [37]. There is also a sharp decrease in the area that receives a certain amount of rainfall and number of rainy days during the season. One study found that the seasonal mean all-India rainfall did not show a significant trend since 1950s [38]. The researchers argued that this is because that the contribution from increasing heavy events is offset by decreasing moderate events. Apparently, should the trend continue, a substantial increase in hazards related to heavy rain would be expected over central India in the future. In another recent study [72], the authors stressed the need to subdivide Indian rainfall geographically and to distinguish early and peak monsoon seasons for the purpose of rainfall trend detection and attribution. They found fingerprints of absorbing aerosols impact on regional rainfall since 1960s featuring increased rainfall in north and northwestern India in May–June and decreased rainfall in central and southern India in July–August since 1960s. However, whether the observed monsoon systems have strengthened or reduced is still very much an open question, limited by the availability of reliable long-term data record. Based on the dramatic decline in ENSO-monsoon correlation in recent decades, it has been suggested that warming over Eurasia continent might have already led to a favored condition for strong monsoons [9]. Besides the potential cause of global warming behind these trends, anthropogenic aerosols could also be a significant factor based on observation-based and modeling studies discussed in previous sections.

Besides anthropogenic influences, natural variability could also be responsible for decadal to centennial variability of the monsoon. Natural variability in the African monsoon over the past three millennia has been reconstructed using geochemical evidence from the sediments of Lake Bosumtwi, Ghana [73]. It was found that intervals of severe drought lasting for periods ranging from decades to centuries are characteristic of the monsoon and are linked to natural variations in Atlantic sea surface temperatures. The researchers thus believe that the severe drought of recent decades is not anomalous in the context of the past three millennia.

The use of climate models to understand causes and consequences of changes in the past and future monsoon climate system is necessary when complex relationships are under consideration. Patricola and Cook used a regional climate model to study the West Africa monsoon in the African Humid Period (AHP; about 14,800–5,500 years ago) when humidity was increased over Africa based on paleoclimate evidence suggesting that the West African summer monsoon was stronger than today, and the Saharan Desert was green [74]. The model was driven by prescribed changes in insolation, atmospheric CO_2, and vegetation to impose conditions at 6,000 years before present, with SSTs fixed at present-day values. The model simulation produced a precipitation increase across the Sahel and Sahara that is in good agreement with the paleoclimate data. They found the precipitation increase in the Sahel is related to a northward shift of monsoon, the elimination of the African easterly jet, and intensification and deepening of the low-level westerly jet on the west coast. Interestingly, the thermal low-Saharan high system of the present-day climate is replaced by a deep thermal low. Even though solar forcing is the ultimate cause of the AHP, the model responded more strongly to the vegetation forcing, emphasizing the importance of vegetation in maintaining the intensified monsoon system.

Takata et al. carried out a pair of climate model simulations using according land use estimates over China and India in 1700 and 1850 [75]. The comparison between these two runs isolates the climate responses to the two different land use estimations. It was found that land use change over China and India from 1700 to 1850 due to population growth (forest to cropland) led to a reduction of surface roughness and thus a weakening of monsoon circulation and precipitation in India.

Various groups have also studied the influence of projected future climate warming on monsoon evolution. Coupled atmosphere-ocean GCM simulations suggested that the increase of surface temperature due to a doubling of CO_2 concentration could enhance mean precipitation of Indian summer monsoon and, as a partial consequence, interannual variability of area-averaged monsoon rainfall [76]. This is believed to be consistent with the observed large variability associated with warm surface temperature.

The performance of the current GCMs in retrospective modeling of monsoon evolution raises issues on both the capability of these models in projecting monsoon in future climate and certain variabilities used to evaluate the models (e.g., the ENSO-Indian summer monsoon rainfall relation). In a recent study, four out of 18 models participated in IPCC AR4 were selected based on their performance of the twentieth-century modeling of monsoon evolution [31]. An analysis was done then for each of these four models using their results from integrations in which the atmospheric CO_2 concentration doubled over preindustrial values. These selected models in the double CO_2 simulations all projected an increase both in the mean monsoon rainfall over the Indian subcontinent (by 5–25%) and in its interannual variability (5–10%). For each model the ENSO-monsoon correlation in the global warming runs is very similar to that in the twentieth-century runs, suggesting that the ENSO-monsoon connection will not weaken as global climate warms. This result is, however, curiously inconsistent with the finding in [9]. In addition, the diversity as seen in the simulations of ENSO variability of these coupled models suggests that these results should be taken with caution.

In a similar study to explore model performance in simulating twentieth-century WAM climatology, Cook and Vizy selected three best-performing models out of

18 coupled GCMs participated in IPCC AR4 to analyze their twenty-first-century integrations under various assumptions about future greenhouse gas increase [22]. Interestingly, each of these three models behaved differently in the twenty-first-century simulations. Only one model projected wet Guinea coast and more frequent dry year in Sahel that are consistent with predicted warming in the Gulf of Guinea based on known dynamic mechanism of the precipitation dipole. The authors thus concluded that there is no consensus among the models concerning the future of the West African monsoon system under greenhouse gas forcing. In another study based on the results from the CMIP3 models, a similar conclusion was obtained that the outlook for Sahel precipitation in these simulations of the twenty-first century is very uncertain, with different models disagreeing even on the sign of the trends [77]. It is especially surprising because most of these models in the twentieth-century integration reproduced the links of Sahel rainfall anomalies to tropical SST anomalies at interannual time scales as shown in observations. Conversely, such a relationship does not explain the rainfall trend in the twenty-first century in a majority of the models.

Future Directions

A key component that limits modeling of monsoon system is observation. The availability of high-resolution surface precipitation data, based on satellite retrievals and surface rain gauge measurements in recent 2 decades, has generally improved the situation. To analyze the longer-term variability, results derived from such dataset need to be compared more carefully with that from the high-density local meteorological measurements. For certain regions such as India, the latter type of data covers more than half century.

Computational technology is advancing rapidly, providing opportunity to model monsoons at higher resolution. However, the speed or memory gain from hardware or software advancements needs to be harnessed not only to increase model resolution but also for improved treatments of physical, chemical, and biogeophysical processes. For modeling the monsoon systems, previous experience suggests that one might need both. One issue that remains a huge challenge in the field of global climate modeling ever since its earliest day is the parameterization of convection. Because the requirement of a-few-kilometer resolution to resolve convection has been a far stretch for GCMs (and will likely still be the case in the near future), description of convection processes for the monsoon system in these models were empirically formulated using parameters resolved in model grid scale. Though still expensive, today's GCMs are already being run in horizontal resolution much closer to the cloud-resolving scale in exploratory and short tests, and perhaps will reach that scale earlier than one would expect. An immediate issue that would come along with this advancement, however, is the realization that variability of monsoon features will also increase at such a high resolution. Clearly, an accurate characterization of monsoon features at the cloud scale exceeds all the current available measurement networks, providing a challenge in the constraint of the high-resolution global climate models by observations.

Instead of using a global high-resolution climate model, the monsoon system can be also simulated in high resolution by coupling the global model with a regional climate model. The latter model usually has a more realistic description of various physical and chemical processes. When needed, running the regional rather than global climate in cloud-resolving scale would greatly reduce the demand for computation. There are numerous attempts reported in literature of so-called downscaling modeling, mostly done by driving a regional model using the output from a global model with a given increment of time (e.g., 6 h). The two-way coupling of such approach, that is, to include certain feedbacks of regional processes to influence the global model, however, is still rare. For modeling the monsoon system, the two-way coupling would provide a better description in the global model of atmospheric feedback to external forcing such as SST anomalies, presuming that the coupled atmosphere and ocean GCMs will remain dominant type of models to cover the global scale.

Studies in recent years have demonstrated the potential influence of aerosols on monsoon circulation and precipitation. Models need to be improved to include more physical- and chemical-based aerosol descriptions, for example, the mixing of anthropogenic aerosol species with dust by the chemical and physical evolution of those particles. The treatment of aerosol–cloud interactions (the so-called indirect effect) is currently quite crude and dependence of these interactions on cloud dynamics is also very crude. Both sets of processes need to be improved in next generation models in order to explore their effects on monsoons. The role of absorbing aerosols revealed in recent studies and limited by measurements of aerosol absorption strength and the distribution of absorbing aerosols also remains a challenge and needs to be much improved.

The onset and evolution of the monsoon system can be influenced by both natural and anthropogenic factors, which are often intertwined both in space and time. Future modeling study should work toward unraveling these two major impacts through better simulation design and advanced statistical analysis. Separating anthropogenic impact from natural variability is essential in narrowing uncertainties in projecting future changes of global climate including monsoon system.

Bibliography

Primary Literature

1. Galvin JFP (2008) The weather and climate of the tropics: part 6 – monsoons. Weather 63:129–137
2. Halley E (1686) A historical account of the trade winds and the monsoon, observable in the seas between and near the tropicks, with an attempt to assign the physical cause of the said winds. Phil Trans R Soc London 16:153–168
3. Webster PJ (1987) The elementary monsoon. In: Fein JS, Stephens PL (eds) Monsoon. Wiley, New York, pp 3–32
4. Krishnamurti TN (1987) Monsoon models. In: Fein JS, Stephens PL (eds) Monsoon. Wiley, New York, pp 467–522
5. Plumb RA (2007) Dynamical constraints on monsoon circulations. In: Schneider T, Sobel AH (eds) The global circulation of the atmosphere. Princeton University Press, Princeton, pp 252–266
6. Blanford HF (1884) On the connexion of the Himalaya snowfall with dry winds and seasons of drought in India. Proc R Soc London 37:3–22
7. Hahn DG, Manabe S (1975) The role of mountains in the South Asian monsoon circulation. J Atmos Sci 32:1515–1541
8. Boos WR, Kuang Z (2009) Dominant control of the South Asian monsoon by orographic insulation versus plateau heating. Nature 463:218–223
9. Kumar KK, Rajagopalan B, Cane MA (1999) On the weakening relationship between the Indian monsoon and ENSO. Science 284:2156–2159
10. Chao WC (2000) Multiple quasi equilibria of the ITCZ and the origin of monsoon onset. J Atmos Sci 57:641–651
11. Chao WC, Chen B (2001) The origin of monsoons. J Atmos Sci 58:3497–3507
12. Prive NC, Plumb RA (2007a) Monsoon dynamics with interactive forcing. Part I: axisymmetric studies. J Atmos Sci 64:1417–1430
13. Prive NC, Plumb RA (2007b) Monsoon dynamics with interactive forcing. Part II: impact of eddies and asymmetric geometries. J Atmos Sci 64:1431–1442
14. Emanuel KA, Neelin JD, Bretherton CS (1994) On large-scale circulations in convecting atmospheres. Quart J R Meteor Soc 120:1111–1143
15. Bordoni B, Schneider T (2008) Monsoons as eddy-mediated regime transitions of the tropical overturning circulation. Nature Geos 1:515–519
16. Wang B, Ding Q, Fu X, Kang I-S, Jin K, Shukla J, Doblas-Reyes F (2005) Fundamental challenge in simulation and prediction of summer monsoon rainfall. Geophys Res Lett 32:L15711. doi:10.1029/2005GL022734
17. Charney JG, Shukla J (1981) Predictability of monsoons. In: Lighthill J, Pearce RP (eds) Monsoon dynamics. Cambridge University Press, New York, pp 99–109
18. Lau KM, Walsier D (2005) Intraseasonal variability in the atmosphere-ocean climate system. Springer Praxis, Chichester, 289 pp
19. Manabe S, Hahn DG, Holloway JL Jr (1974) The seasonal variation of the tropical circulation as simulated by a global model of the atmosphere. J Atmos Sci 31:43–83
20. Gates WL et al (1999) An overview of the results of the atmospheric model intercomparison project (AMIP I). Bull Am Meteorol Soc 80:29–56
21. Shen B-W, Tao W-K, Lau WK, Atlas R (2010) Improving tropical cyclogenesis prediction with a global mesoscale model: hierarchical multiscale interactions during the formation of tropical cyclone nargis (2008). J Geophys Res 115:D14102. doi:10.1029/2009JD013140
22. Cook KH, Vizy EK (2006) Coupled model simulations of the West African monsoon system: twentieth- and twenty-first-century simulations. J Clim 19:3681–3703
23. Krishnamurthy V, Shukla J (2007) Intraseasonal and seasonally persisting patterns of Indian monsoon rainfall. J Clim 20:3–20
24. Shukla J (1975) Effect of Arabian sea-surface temperature anomaly on Indian summer monsoon: a numerical experiment with the GFDL model. J Atmos Sci 32:503–511
25. Lau K-M, Yang S (1996) Seasonal variation, abrupt transition, and intraseasonal variability associated with the Asian summer monsoon in the GLA GCM. J Clim 9:965–985
26. Meehl GA, Arblaster JM, Lawrence D, Seth A, Schneider EK, Kirtman BP, Min D (2006) Monsoon regimes in the CCSM3. J Clim 19:2482–2495
27. Christensen JH, Hewitson B, Busuioc A, Chen A, Gao X, Held I, Jones R, Kolli RK, Kwon W-T, Laprise R, Magaña Rueda V, Mearns L, Menéndez CG, Räisänen J, Rinke A, Sarr A, Whetton P (2007) Regional climate projections, in climate change 2007: the physical science basis. In: Solomon S, Qin D, Manning M, Chen Z, Marquis M,

Averyt KB, Tignor M, Miller HL (eds) Contribution of working group I to the fourth assessment report of the intergovernmental panel on climate change. Cambridge University Press, New York

28. Meehl GA (1994) Coupled land-ocean-atmosphere processes and South Asian monsoon variability. Science 266:263–267
29. Meehl GA, Arblaster JM (2002) Indian monsoon GCM sensitivity experiments testing tropospheric biennial oscillation transition conditions. J Clim 15:923–944
30. Shukla J (2007) Monsoon mysteries. Science 318:204–205
31. Annamalai H, Hamilton K, Sperber KR (2007) The South Asian summer monsoon and its relationship with ENSO in the IPCC AR4 simulations. J Clim 20:1071–1092
32. Xue YK et al. (2010) Intercomparision and analyses of the West African monsoon in the West African monsoon model evaluation (WAMME) project: first model intercomparison experiment. Climate Dynamics 35:3–27. doi:10.1007/s00382-010-0778-2
33. Giannini A, Saravanan R, Chang P (2003) Oceanic forcing of Sahel rainfall on interannual to interdecadal time scales. Science 302:1027–1030
34. Lau KM, Kim MK, Kim KM (2006) Asian monsoon anomalies induced by aerosol direct effects. Clim Dyn 26:855–864. doi:10.1007/s00382-006-0114-z
35. Lu R, Fu Y (2010) Intensification of East Asian summer rainfall interannual variability in the twenty-first century simulated by 12 CMIP3 coupled models. J Clim 23:3316–3331
36. Ramanathan V, Chung C, Kim D, Bettge T, Buja L, Kiehl JT, Washington WM, Fu Q, Sikka DR, Wild M (2005) Atmospheric brown clouds: impact on South Asian climate and hydrologic cycle. Proc Natl Acad Sci USA 102:5326–5333
37. Ramesh KV, Goswami P (2007) Reduction in temporal and spatial extent of the Indian summer monsoon. Geophys Res Lett 34:L23704. doi:10.1029/2007GL031613
38. Goswami BN, Venugopal V, Sengupta D, Madhusoodanan MS, Xavier PK (2006) Increasing trend of extreme rain events over India in a warming environment. Science 314:1442–1445
39. Wang C (2004) A modeling study on the climate impacts of black carbon aerosols. J Geophys Res 109:D03106. doi:10.1029/2003JD004084
40. Roberts DL, Jones A (2004) Climate sensitivity to black carbon aerosol from fossil fuel combustion. J Geophys Res 109: D16202. doi:10.1029/2004JD004676
41. Chung SH, Seinfeld JH (2005) Climate response of direct radiative forcing of anthropogenic black carbon. J Geophys Res 110:D11102. doi:10.1029/2004JD005441
42. Lau K-M, Ramanathan V, Wu G-X, Li Z, Tray SC, Hsu C, Sikka R, Holben B, Lu D, Tartari G, Chin M, Koudelova P, Chen H, Ma Y, Huang J, Taniguchi K, Zhang R (2008) The joint aerosol-monsoon experiment, a new challenge for monsoon climate research. Bull Am Meteorol Soc 89:369–383
43. Menon S, Hansen J, Nazarenko L, Luo Y (2002) Climate effects of black carbon aerosols in China and India. Science 297:2250–2253
44. Wang C (2007) Impact of direct radiative forcing of black carbon aerosols on tropical convective precipitation. Geophys Res Lett 34:L05709. doi:10.1029/2006GL028416
45. Randles CA, Ramaswamy V (2008) Absorbing aerosols over Asia: a geophysical fluid dynamics laboratory general circulation model sensitivity study of model response to aerosol optical depth and aerosol absorption. J Geophys Res 113: D21203. doi:10.1029/2008JD010140
46. Meehl GA, Arblaster JM, Collins WD (2008) Effects of black carbon aerosols on the Indian monsoon. J Clim 21:2869–2882. doi:10.1175/2007JCLI1777.1
47. Lau K-M, Kim K-M (2006) Observational relationships between aerosol and Asian monsoon rainfall, and circulation. Geophys Res Lett 33:L21810. doi:10.1029/2006GL027546
48. Bollasina M, Nigam S, Lau K-M (2008) Absorbing aerosols and summer monsoon evolution over South Asia: an observational portrayal. J Clim 21:3221–3239
49. Gautam R, Hsu NC, Lau K-M, Tsay S-C, Kafatos M (2009) Enhanced pre-monsoon warming over the Himalayan-Gangetic region from 1979 to 2007. Geophys Res Lett 36: L07704. doi:10.1029/2009GL037641
50. Wang C, Kim D, Ekman AML, Barth MC, Rasch PJ (2009) Impact of anthropogenic aerosols on Indian summer monsoon. Geophys Res Lett 36:L21704. doi:10.1029/2009GL040114
51. Carlson TN, Prospero JM (1972) The large-scale movement of Saharan air outbreaks over the Northern equatorial Atlantic. J Appl Meteor 11:283–297
52. Prospero JM, Lamb JP (2003) African droughts and dust transport to the Caribbean: climate change and implications. Science 302:1024–1027
53. Dwyer E, Pinnock S, Gregoire JM, Pereira JMC (2000) Global spatial and temporal distribution of vegetation fire as determined from satellite observations. Int J Rem Sens 21:1289–1302
54. Duncan BN, Martin RV, Staudt AC, Yevich R, Logan JA (2003) Interannual and seasonal variability of biomass burning emissions constrained by satellite observations. J Geophys Res 108:4100. doi:10.1029/2002JD002378
55. Ito A, Penner JE (2005) Historical emissions of carbonaceous aerosols from biomass and fossil fuel burning for the period 1870–2000. Global Biogeochem Cy 19:GB2028. doi:10.1029/2004GB002374
56. Huang J, Zhang C, Prospero JM (2009) African dust outbreaks: a satellite perspective of temporal and spatial variability over the tropical Atlantic ocean. J Geophys Res 115:D05202. doi:10.1029/2009JD012516
57. Huang J, Zhang C, Prospero JM (2009) Large-scale effects of aerosols on precipitation in the West African monsoon region. Q J R Meteorol Soc 135:581–594
58. Huang J, Zhang C, Prospero JM (2009) African aerosol and large-scale precipitation variability over West Africa. Environ Res Lett 4:015006. doi:10.1088/1748-9326/4/1/015006
59. Huang J, Adams A, Wang C, Zhang C (2009) Black carbon and West African monsoon precipitation: observations and simulations. Ann Geophys 27:4171–4181

60. Eltahir E, Gong C (1996) Dynamics of wet and dry years in West Africa. J Climate 9:1030–1042
61. Lau KM, Kim KM, Sud YC, Walker GK (2009) A GCM study of the response of the atmospheric water cycle of West Africa and the Atlantic to Saharan dust radiative forcing. Ann Geophys 27:4023–4037, http://www.ann-geophys.net/27/4023/2009/
62. Kim KM, Lau KM, Sud YC, Walker GK (2010) Influence of aerosol-radiative forcing on the diurnal and seasonal cycles of rainfall over West Africa and the Eastern Atlantic using GCM simulations. Clim Dyn. doi:10.1007/s00382-010-0750-1
63. Wilcox E, Lau WKM, Kim KM (2010) A northward shift of the Inter-tropical convergence zone in response to summertime Saharan dust outbreak. Geophys Res Lett 37:L04804. doi:10.1029/2009GL041774
64. Rotstayn LD, Cai W, Dix MR, Farquhar GD, Feng Y, Ginoux P, Herzog M, Ito A, Penner JE, Roderick ML, Wang M (2007) Have Australian rainfall and cloudiness increased due to the remote effects of Asian anthropogenic aerosols? J Geophys Res 112: D09202. doi:10.1029/2006JD007712
65. Cai W, Cowan T, Sullivan A, Ribbe J, Shi G (2011) Are anthropogenic aerosols responsible for the Northwest Australia summer rainfall increase? A CMIP3 perspective and implications. J Climate 24:2556–2564. doi:10.1175/2010JCLI3832.1
66. Wang Y, Cheng H, Edwards RL, He Y, Kong X, An Z, Wu J, Kelly MJ, Dykoski CA, Li X (2005) The Holocene Asian monsoon: links to solar changes and North Atlantic climate. Science 308:854–857
67. Wang Y, Cheng H, Edwards RL, Kong X, Shao X, Chen S, Wu J, Jiang X, Wang X, An Z (2008) Millennial- and orbital-scale changes in the East Asian monsoon over the past 224,000 years. Nature 451:1090–1093
68. Zhang P, Cheng H, Edwards RL, Chen F, Wang Y, Yang X, Liu J, Tan M, Wang X, Liu J, An C, Dai Z, Zhou J, Zhang D, Jia J, Jin L, Johnson KR (2008) A test of climate, sun, and culture relationships from an 1810-year Chinese cave record. Science 322:940–942
69. Anderson DM, Overpeck JT, Gupta AK (2002) Increase in the Asian Southwest monsoon during the past four centuries. Science 297:596–599
70. Goes JI, Thoppil PG, Gomes HDR, Fasullo JT (2005) Warming of the Eurasian landmass is making the Arabian Sea more productive. Science 308:545–547
71. Dash SK, Kulkarni MA, Mohanty UC, Prasad K (2009) Changes in the characteristics of rain events in India. J Geophys Res 114: D10109. doi:10.1029/2008JD010572
72. Lau KM, Kim K-M (2010) Fingerprinting the impacts of aerosols on long-term trends of the Indian summer monsoon regional rainfall. Geophys Res Lett 37:L16705. doi:10.1029/2010GL043255
73. Shanahan TM, Overpeck JT, Anchukaitis KJ, Beck JW, Cole JE, Dettman DL, Peck JA, Scholz CA, King JW (2009) Atlantic forcing of persistent drought in West Africa. Science 324:377–380
74. Patricola CM, Cook KH (2007) Dynamics of the West African monsoon under mid-Holocene precessional forcing: regional climate model simulations. J Clim 20:694–716
75. Takata K, Saito K, Yasunari T (2009) Changes in the Asian monsoon climate during 1700–1850 induced by preindustrial cultivation. Proc Natl Acad Sci USA 106:9586–9589
76. Meehl GA, Washington WM (1993) South Asian summer monsoon variability in a model with doubled atmospheric carbon dioxide concentration. Science 260:1101–1104
77. Biasutti M, Held IM, Sobel AH, Giannini A (2008) SST forcings and Sahel rainfall variability in simulations of the twentieth and twenty-first centuries. J Clim 21:3471–3486

Books and Reviews

Fein JS, Stephens PL (1987) Monsoon. Wiley, New York

Lighthill J, Pearce RP (eds) (1981) Monsoon dynamics. Cambridge University Press, New York

Wang B (2006) East Asian monsoon. Springer/Praxis, Chichester

Webster PJ, Magana VO, Palme TN, Shukla J, Tomas RA, Yanai M, Yasunari T (1998) Monsoons: processes, predictability, and the prospects for prediction. J Geophys Res 103:14451–14510

Mussel Culture, Open Ocean Innovations

Richard Langan
Atlantic Mariner Aquaculture Center, University of New Hampshire, Durham, NH, USA

Article Outline

Glossary

Suspension culture A production method for mussels and other shellfish that employs ropes, cages, or nets suspended in the water column from either rafts or longlines.

Surface longline An anchored structure consisting of surface floatation supporting one or more horizontal lines from which ropes, cages, or nets can be suspended in the water column.

Open ocean farming Refers to aquaculture production of marine organisms in open ocean or offshore waters that are removed from any significant influence of land masses.

Submerged longline Subsurface structure consisting of anchors and submerged floatation from which ropes, cages, or nets can be suspended.

Site selection The process for selecting farming sites based on specified parameters such as depth, current and wave climate, temperature, and primary productivity.

Environmental effects The effects of farming activities on the physical, biological, and chemical properties of the marine environment *and* the effects of the environment on cultured organisms and consumers of cultured food products.

Seston Particulate material suspended in the water column of water bodies consisting of both living and dead organic material and inorganic particles.

Pseudofeces Suspended particles that have been rejected as food by filter feeding bivalve mollusks. The rejected particles are wrapped in mucus and expelled without being passed through the digestive tract.

Definition of the Subject

Aquaculture production of several species of mussels in sheltered marine waters is well established and occurs in many countries worldwide. The primary method of production of high quality mussels is suspension of ropes with attached mussels from floating rafts or surface longlines that are anchored to the seafloor. While demand for fresh, frozen, and canned mussel products continues to increase, growth in production is hampered by a lack of suitable space for expansion in sheltered waters. For more than a decade, there has been interest in developing production methods suitable for open ocean environments where wind and wave conditions preclude the use of either rafts or surface longlines. Recent advances in the use of longlines that can be submerged below the sea surface and therefore avoid the upper portion of the water column that is most affected by wave energy indicate that open ocean production is feasible. However, additional development in technology and methods to improve production efficiency and insure worker safety, as well as changes to political and regulatory frameworks are needed in order to achieve large-scale production.

Introduction

Population growth and consumer preference have resulted in a growing demand for seafood, a trend that is projected to continue into the future [1]. Production from capture fisheries has leveled off, and by most projections will remain stagnant or decline, depending on management and regulatory measures implemented by fishing nations [2, 3]. In contrast, aquaculture production has increased by nearly 10% each year since 1980, and has played an important role in filling the gap between seafood supply and demand. Only a few decades ago, wild-caught fish and shellfish supplied nearly all edible seafood, though with essentially flat growth since 1980 and the rise of aquaculture over the same time period, capture fishing now accounts for only about half of the total [1]. In the most optimistic scenarios, wild-caught fisheries production will remain stagnant [2]; therefore, growth in the global seafood supply will continue to rely on aquaculture production.

There are signs, however, that the rate of growth for global aquaculture may have peaked for land-based and nearshore marine culture due to political, environmental, economic, and resource constraints [1]. Expansion of land-based culture is limited primarily by economics, particularly in developed countries where costs associated with land, capital equipment, and energy required to pump and treat water are prohibitive. In addition, very few marine species are appropriate for land-based culture. For example, the space and volume of phytoplankton required to produce large quantities of filter feeding mollusks in land-based systems would be enormous, and therefore not economically viable.

For nearshore marine farming, available and suitable space is the primary limiting factor as sheltered coastal waters are for most countries quite constrained to begin with and are already used for a multitude of commercial and recreational activities with which aquaculture must compete for space [4]. Expansion of large-scale finfish farming in coastal waters is also limited by environmental concerns. While there are also concerns about potential environmental effects of bivalve mollusk culture, they are minor in comparison to net pen culture of finfish and are balanced by recognition of the ecosystem services such as enhanced habitat complexity and filtration capacity provided by mollusks [5]. It is rather the effect of environmental conditions on mollusk culture, and specifically the effects of pollution on product safety that is limiting expansion in nearshore waters. Rapid coastal development and population growth and the resulting increase in human sources of pollution have affected the sanitary quality of nearshore waters, rendering shellfish grown there unsafe for consumption. As a consequence, many otherwise suitable sheltered sites for mollusk culture are off limits due to public health restrictions.

In developed countries, conflict with coastal residents and tourist-related businesses over aesthetic values, primarily over water views from shorefront property, have also affected the establishment of new farming sites. As the demographic of coastal communities continues to change and new residents place more value on views and recreation than food production, these conflicts are likely to increase. Given the constraints on expansion of current methods of production, it is clear that alternative approaches are needed in order for the marine aquaculture sector to make a meaningful contribution to the world's seafood supply.

Farming in open ocean waters has been identified as one potential option for increasing production and has been a focus of international attention for more than a decade. Despite the global interest in open ocean farming, development to date has been measured, primarily due to the significant technical and operational challenges posed by wind and wave conditions in most of the world's oceans [4]. Farming in fully exposed open ocean waters requires a different engineering approach since equipment and methods currently used in sheltered nearshore sites are largely unsuitable for the open ocean. In addition, the scale of investment required to develop and demonstrate new technologies and methods for offshore farming is yet to be determined, though most engaged in this endeavor would agree that it will likely be substantial.

Despite these challenges, there is sufficient rationale for pursuing the development of open ocean farming. Favorable features of open ocean waters include ample space for expansion, tremendous carrying capacity, less conflict with many user groups, reduced exposure to human sources of pollution, the potential to moderate some of the negative environmental and aesthetic impacts of high density coastal farming [6–8], and optimal environmental conditions for some bivalve mollusk species [9, 10]. For many countries, where cost, environmental concerns, limited space, and competing uses have restricted growth of land-based and nearshore marine farming, few other options for significant expansion exist.

Of the many species of finfish and shellfish that have been considered for open ocean farming, several species of mussels have emerged as attractive candidates. There are several reasons for this. Like all filter-feeding mollusks, mussels derive all their nutritional needs from naturally occurring phytoplankton and organic particulates. Therefore, daily visits to deliver formulated feed by service vessels and farm personnel, which may be prohibited for extended periods by sea conditions, are not needed, nor is on-site infrastructure for automated feeding, which is both costly and vulnerable to damage from storms. Unlike many cultured species that have gradually transitioned from wild capture to aquaculture, farming has been the primary means of production for mussels for many decades; therefore, methods used in sheltered waters are well developed, highly automated, and very efficient [11]. Mussels are also relatively fast growers, with production cycles ranging from 12 to 18 months [9, 12].

Production methods in sheltered nearshore waters include bottom culture, which is practiced in some locations such as the Netherlands, Scandinavia, and the USA (Maine), and pole or "bouchot" culture, which is practiced in France; however, suspension culture, because of superior product quality, accelerated growth, and opportunities for mechanization, has emerged as the leading method of production [11]. Techniques and materials used for suspension culture may vary somewhat from place to place; however, in general, culture methodology consists of suspending mussel ropes or "droppers" from either rafts or longlines [13]. Raft culture was pioneered in Spain and from there became established in Scotland and more recently in Maine USA and in the Pacific Northwest coast of North America [11]. While rafts can be highly productive, they are suitable for use only in very sheltered embayments. Longline technology, which was developed in Japan, consists of either surface or submerged longlines, held in place with anchors and supported by buoys or floats. As with raft culture, surface longlines are only suitable for use in sheltered waters [13]; therefore, in locations where adverse sea conditions or drift ice occur, submerged longlines are the only option. Submerged longlines have been used primarily in locations (e.g., Atlantic Canada) where winter ice would impact buoys and lines [14]. It is only in recent years that the technology has been used in fully exposed open ocean locations [9].

Characterization and Selection of Open Ocean Farming Sites

Before discussing approaches to the development of open ocean mussel culture, it is important to first define what is meant by the term "open ocean." For most engaged in this sector, it is used synonymously with "offshore" and is generally accepted to mean farming in locations that are subjected to ocean waves and currents and removed from any significant influence of land masses rather than a set distance from shore. Clearly, a wide range of sea conditions falls under this broad definition. Ryan [4] reported on a site classification system for marine waters developed in Norway that is based on significant wave height exposure (Table 1).

While this classification method is instructive, knowledge of the full range of conditions at a particular site is needed to develop appropriate technologies and safe and efficient operating procedures.

Mussel Culture, Open Ocean Innovations. Table 1 Norwegian classification of offshore waters based on significant wave heights (From Ryan [4])

Site Class	Significant wave height (m)	Degree of exposure
1	<0.5	Small
2	0.5–1.0	Moderate
3	1.0–2.0	Medium
4	2.0–3.0	High
5	>3.0	Extreme

There are a number of criteria that determine the suitability of open ocean sites for farming, many of which are also considerations for sheltered waters. These include proximity to infrastructure such as ports, processing and distribution centers, as well as physical and biological criteria such as bathymetry, seabed characteristics and contour, current velocities, temperature profiles, dissolved oxygen, turbidity, the quantity of quality of phytoplankton, and the frequency of occurrence of harmful algal blooms. The most important additional feature of offshore sites is wave climate. Significant wave heights, wave periods, the frequency and duration of high energy storm conditions, and the combined forcing of waves and currents must be known in order to determine whether a site is suitable, accessible by service vessels and personnel with reasonable frequency, and if so, what type of technology is required for farming.

It is imperative that a thorough evaluation of the parameters described above be conducted before proceeding with development of a site for farming. The requirements for data and subsequent analysis can be substantial; however, the use of advanced oceanographic technologies can greatly facilitate this task [8]. Multibeam sonar and three-dimensional visualization can generate a wealth of data on seafloor contours and texture to inform mooring system design and placement. Collection of time intensive data on temperature, salinity, dissolved oxygen, turbidity, and fluorescence can be greatly facilitated by strategic deployment of in situ instrumentation at appropriate depth intervals in the water column. Additional instrumentation should include Acoustic Doppler Current Profilers (ADCP) that can measure and record current velocity and direction throughout the water column, wave sensors that can give precise data on wave height, direction, steepness, and period, and meteorological sensors to measure air temperature and wind speed and direction. Many countries have buoy arrays in coastal waters that can provide long-term data on regional climatology to aid site evaluation; however, collection of site-specific data is critical. Assessment of the potential for the effects of global climate change on critical parameters such as water temperature should also be considered.

The data collection period required for site evaluation will vary, depending on local and regional environmental and meteorological conditions. Good baselines for some parameters can be established in a relatively short time frame (1 year), others such as the frequency, duration, and severity of storms or blooms of toxic algae are less predictable and it may take longer to determine the suitability of a particular site.

While most of the focus on open ocean development has been on cage culture of finfish, there has also been growing interest in offshore culture of bivalve mollusks. Some of the same drivers such as ample space and the opportunity to avoid user conflicts are identical to those for finfish culture, though perhaps more importantly, reduced risk of exposure to human sewage and industrial pollution presents a major advantage of open ocean waters over coastal locations.

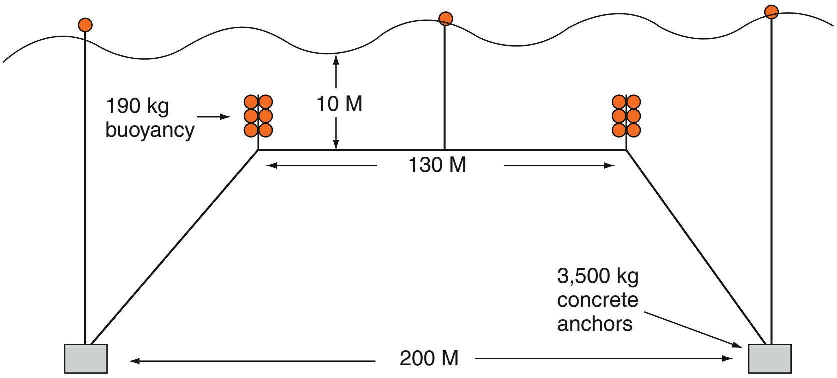
There are, however, possible limitations as well as advantages. Open ocean waters in many areas of the world are nutrient deficient, so careful attention must be paid during site selection to the quantity, quality, and seasonality of phytoplankton available to dense arrays of filter feeding mollusks. Macroscale information on primary productivity can be obtained from ocean color satellite data generated by instruments such as Sea-viewing Wide Field-of-view Sensor (SeaWiFS) and Moderate Resolution Imaging Spectroradiometer (MODIS). Site-specific data on concentration and composition can be generated by in situ fluorometry and microscopic analysis of the plankton community. Phytoplankton concentration at different depths is also an important factor, as farmers will wish to maximize the use of vertical space for production in deep ocean waters. The frequency and duration of harmful algal blooms (HABs) is also a critical consideration for offshore mollusk farming. In some locations, blooms of toxic algae originate and persist in offshore waters (e.g., *Alexandrium sp.* In the Gulf of Maine, USA) and can result in extended public health closures with severe economic impact on producers.

In addition to physical, chemical, and biological characteristics of a site, other human uses in the vicinity such as shipping, fishing, and mining must be identified in order to avoid conflicts. Involvement of the appropriate permitting authorities in the early stages of development of an open ocean farming site is also critical [15]. Other factors such as use of the area by marine mammals, proximity to foraging areas of predators (e.g., diving ducks), location of sensitive biological communities, presence of parasitic organisms (e.g., pea crabs, trematodes, and copepods), and sediments contaminated by toxic substances must also be considered [16].

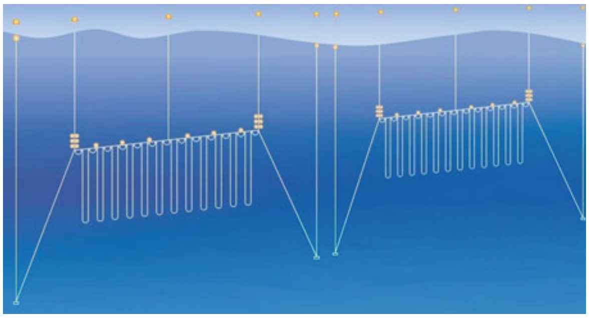
Technologies for and Methods Open Ocean Mussel Farming

Technologies for open ocean mussel farming are essentially adaptations of suspension culture methods employed in sheltered marine waters. Designs and prototypes for submersible rafts have been developed [17, 18]; however, submerged longlines are the most commonly used method. This technology was developed in Japan and has been in use there for several decades for deep water suspended scallop culture, though not in fully exposed open ocean conditions. The technology has been successfully adapted for sheltered water mussel culture in Atlantic Canada where winter and spring drift ice can damage surface longlines [14]. More recently, the technology has been shown to be effective for mussel production in very high-energy open ocean conditions (e.g., significant wave heights >10 m) in the northeast USA [9] and at a test site in the German Bight with significant wave heights >8 m, and current velocity up to 1 ms^{-1} [19]. The technology is quite simple and it consists of relatively inexpensive materials. A design currently in use in North America is presented in Fig. 1.

The structural stability of a submerged longline is maintained by the opposing forces of submerged flotation at the ends of a single horizontal backbone, connected by lines set at a 45° angle to seafloor anchors. The most commonly used anchors are large (3–6 tons) deadweight concrete anchors, though both plow type and screw anchors have been used in some locations. Submergence depth of the backbone is dictated by site-specific wave climate and can range from 3 to 15 m. Surface floatation is minimized to prevent the transfer of wave-induced motion the backbone, and consists of nonstructural marker buoys for the anchor lines and a mid-backbone pick-up line that provides access to the crop from a service vessel. Anchors are generally spaced from 100 to 200 m apart, and depending upon the



Mussel Culture, Open Ocean Innovations. Figure 1
A schematic of a submerged longline used for suspension culture of mollusks in open ocean environments

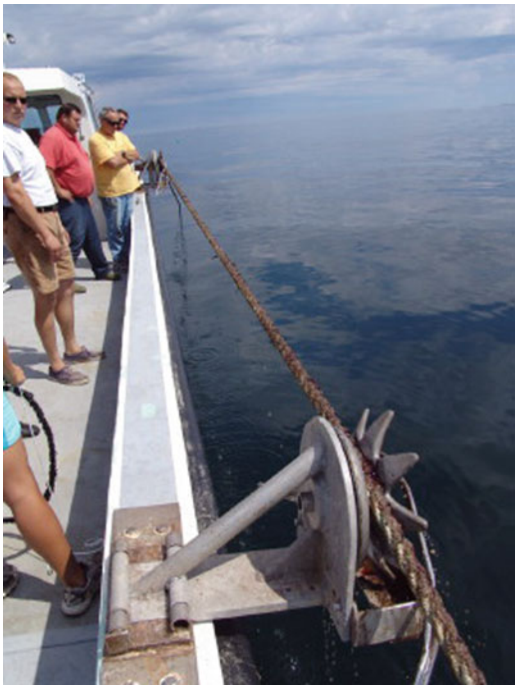


Mussel Culture, Open Ocean Innovations. Figure 2
A diagram of a submerged longlines showing the attachment of mussel growing ropes to the backbone and the placement of floatation added to the backbone as the crop increases in mass during growout (From Langan and Horton [9])

depth of the water and desired depth of submergence, the backbone length can range from 70 to 130 m. Ropes or “droppers” of mussels are suspended from the backbone, and additional submerged floatation is added as the crop gains mass during growout (Fig. 2).

At some of the open ocean farms that have been established, converted fishing vessels are currently used to tend offshore longlines. The deck equipment required for tending lines to seed growout ropes and to inspect and harvest crops is similar to that in use for sheltered sites and includes rail mounted starwheels (Fig. 3) and an articulating crane (Fig. 4).

In addition, equipment common to many fishing vessels such as a lobster or crab trap hauler or a rotating boom is needed for lifting the submerged line to the surface. If there is sufficient deck space, bulk processing equipment such as declumping and debyssing machines can be used during harvest operations to reduce the need for extensive processing at shore-based



Mussel Culture, Open Ocean Innovations. Figure 3
A forward looking view of the starboard side of a service vessel showing the backbone of a submerged longline set into aft (foreground) and forward starwheels. Growing ropes with seed mussels are attached to the backbone for the growout cycle

Mussel Culture, Open Ocean Innovations. Figure 4 A hydraulic articulating crane on a service vessel, shown here being used to unload equipment, is used extensively in mussel farming operations

facilities. Though converted fishing vessels may be used as this sector develops, it is likely that large, seaworthy, specialized vessels that can carry the harvesting and primary processing gear, provide a stable platform for lifting operations and a large load capacity for the harvest will be required to support large-scale operations. Vessels of this nature are in use in France and New Zealand [20].

In addition to submerged longlines, some experimental efforts have employed a submersible ring-like structure attached to a wind turbine tower, which has been used for offshore macroalgae growout [21]. This device could potentially be used for mussel cultivation; however, there may be scaling issues in reaching the desired biomass.

Mussel Species in Open Ocean Cultivation

There are several species of mussels that are cultivated in open ocean waters; however, regardless of species or location, production is currently minor by comparison with well-established nearshore production sites. In North America, small quantities of blue mussels (*Mytilus edulis)* are produced in offshore farms in New England (USA) and Atlantic Canada and Mediterranean mussels (*M. galloprovincialis*) are being grown at an offshore farm off the southern California (USA) coast [22]. In Europe, *M. galloprovincialis* are grown on submerged longlines at exposed locations in the Mediterranean coast of France [23] and in the Turkish Black Sea. Culture trials have been initiated for *M. edulis* in the North Sea off the coast of Germany, [19] and in the Belgian North Sea [24]. Other European countries, including Portugal, Spain, Italy, and Ireland are developing strategies for offshore mussel production.

In New Zealand, where the nearshore greenshell mussel (*Perna canaliculus*) industry is well developed and highly mechanized, there is a great deal of interest in developing large-scale ocean farms, as lease sites in sheltered nearshore waters have become difficult to obtain [25]. Initial efforts at open ocean mussel farming involved moving the double longline surface technology into more exposed sites and some success was achieved in wave conditions up to 2.5 m [26]. However, failure of surface longline systems in higher energy sites has led to the development of submerged technologies and a small number of open ocean mussel farms are operating in New Zealand offshore waters, with many new farms proposed [27]. This scale of expansion is projected to provide a threefold increase in production and export earnings by 2020 [28].

While data is limited to a few locations in North America and France, there are indications that production cycles and product quality for mussels grown in open ocean waters are highly favorable. Open ocean farms off the New Hampshire coast in the northeast USA have consistently produced market-sized (55 mm) blue mussels in 12–14 months from spat settlement with meat yields ranging from 42% to 58% [9]. Similar data has been reported for blue mussels at sites off the coast of Martha's Vineyard [29]. By comparison, rope-grown blue mussels from nearby estuaries and bays can take up to 18 months to reach market size [30]. Mediterranean mussels produced at an open ocean site in California have also demonstrated excellent growth and quality, reaching market size in 6–8 months and nearly 50% meat yield [22]. Trials in the North Sea have shown that the growth conditions in the German Bight are very favorable for mussel cultivation. Market-size (50–55 mm) can be reached by 12–15 months and infestation by parasites is much lower than in near-shore sites [10]. Faster growth at offshore sites may to be due to a more stable temperature and salinity conditions and therefore lower stress, reduced turbidity, and better water exchange [20].

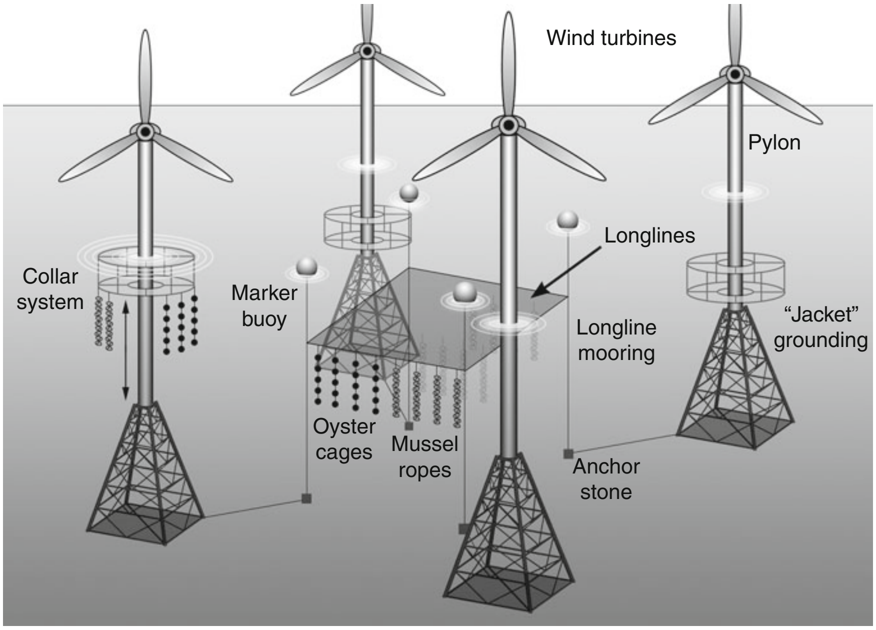
Open Ocean Mussel Farming in Multiuse Facilities

Open ocean mussel farming can be practiced in isolation of other activities; however, there may be economic or environmental advantages to combining mussel culture with offshore fish farming or energy production. At a nearshore marine farming site in New Brunswick, Canada, Lander et al. [31] demonstrated better growth rates for raft-cultured mussels 100 m down current of a salmon farm than at reference sites, and was able to document that organic wastes, primarily fine particulates from feed emanating from the salmon farm contributed to the diet of the mussels. In open ocean sites, creating mussel culture "zones" in proximity to finfish farms may offset the effects of organic loading to the environment [32].

Energy installations may also provide structure for deployment of mussel culture systems. Mussels (*M. galloprovincialis*) have been harvested from oil platforms in California, USA for many years [33], and there is interest in using decommissioned offshore oil platforms as attachment points for mussel culture infrastructure.

Buck et al. [34] investigated the possibility of integrating suspension culture of oysters and mussels at existing offshore wind energy platforms in the North Sea (Fig. 5).

There are a number of advantages for conducting mussel cultivation activities within the footprint offshore wind farms. The placement of aquaculture production facilities in defined corridors between wind farm turbines eliminates the need for a separately permitted facility and reduces the space required if the two facilities were located separately [34]. Also, infrastructure for regular servicing may be shared. As both industries need a multifunctional service vessel, preferably with lifting capacities to install and change plant components and execute farming operations, and sufficient deck space to carry equipment and stock, the opportunity to share high-priced infrastructure exists [35].



Mussel Culture, Open Ocean Innovations. Figure 5
A schematic of shellfish growing systems associated with wind turbine towers (from Buck et al. [34])

Further, a combined environmental impact assessment for both users may reduce costs.

Environmental Considerations for Open Ocean Mussel Farming

Like all forms of food production, the culture of marine species, whether practiced in land-based, nearshore, or open ocean locations will have some effect on the environment. The effect can be both negative and positive and can vary depending upon the species, location, and farming practices. In the past 3 decades of marine farming in sheltered marine waters, adverse impacts from aquaculture of both molluskan shellfish and finfish have been documented, though most of the concerns and controversy are centered on finfish. Mollusk culture is generally perceived as environmentally benign or even beneficial [5]; however, there have been documented environmental impacts from nearshore mussel farming that merit consideration for development of the offshore sector.

Though mussels feed on naturally occurring seston and no external feed is provided to the organisms, deposition of feces and pseudofeces can enrich bottom sediments beneath culture systems and impact benthic communities [36, 37]. Occurrences of sediment impacts have been associated with very dense culture in shallow embayments; therefore, if offshore farms are sited in locations with sufficient depth and adequate water circulation to disperse wastes, enrichment of bottom sediments should not be an issue [7]. High-density mussel culture can also deplete the water column of planktonic food, affecting both the growth and fitness of the cultured organisms as well as naturally occurring filter feeders in the system [38]. This too, is an impact that has been observed in sheltered embayments with limited circulation and is unlikely to be an environmental issue in open ocean waters [8]. However, in very large, high-density offshore farms, depletion of food within the farm and reduced growth and condition of the stock may be an issue for producers.

Hydrodynamic alteration is another environmental effect that has been documented in sheltered embayments with high-density shellfish culture [39] and has recently been an issue of concern in New Zealand where large-scale open ocean mussel farming is in development. Plew et al. [28] reported significant current and wave attenuation and strong water column stratification at a large (230 longline) mussel farm in Golden Bay, New Zealand. The farm was located in relatively shallow water (10–12 m) and the culture organisms were suspended from the surface to a depth of 8 m, therefore, occupying nearly the entire water column. As it is likely that open ocean development will use submerged culture in much deeper water (30–100 m) with ample space above and below the culture arrays, the severity of flow modifications as observed in this study are improbable.

A legitimate environmental concern for open ocean mussel culture is entanglement of whales and other marine life in seed collection lines [40]. These collectors are either discrete lengths of line or one continuous length of rope suspended from the backbone to provide substrate for settlement of mussel larvae (Fig. 6). As this sector develops, it is important to avoid deployment of seed collection lines in the migratory pathways of endangered marine mammals or to use weak links and electronic alert systems in the farming infrastructure [41].

Mussel Culture, Open Ocean Innovations. Figure 6
Seed collecting rope (*black*) is attached to the backbone of a submerged longline

Future Directions

Developments over the past 2 decades indicate that aquaculture production of mussels in open ocean environments is feasible and that opportunities exist for large-scale production [9, 10]. Conflicts with other uses can be significantly reduced, though they are not totally eliminated [34]. There is also evidence to support the premise that environmental impacts can be reduced by farming in open ocean environments [8, 36]. There is also strong indication that if sites are chosen properly, faster growth and excellent product quality can be achieved [9].

Though some technical challenges remain such as the development of large, purpose built, and highly seaworthy service vessels, obstacles to development of open ocean mussel farming are primarily economic, social, and political in nature. The scale of investment needed to establish and operate large-scale open ocean mussel farms is not well known, though it is assumed that production costs will be higher than for nearshore farming. The additional costs could be partially offset if ocean grown mussels, due to superior quality and greater consumer confidence in product safety can command a higher price [9], however, market prices are subjected to many economic externalities that are difficult to forecast. Space conflicts with the fishing industry may be an issue in some locations, therefore, involvement of local capture fishermen in industry development may be needed to gain acceptance of an alternative use of ocean space. As many countries move toward spatial planning of their territorial ocean waters, it is important to include a future vision of the potential for open ocean mussel farming in the planning process and give due consideration to compatibilities and possible synergies with other uses. Many countries also currently lack the regulatory framework for permitting open ocean farming sites. Until economic and regulatory uncertainties are resolved, entrepreneurs will be reluctant to make the level of investment needed to move this sector forward.

Ideally, development of open ocean farming should take place within the context of overall ocean management and marine spatial planning in order to assure compatibility with other uses and consistency with broader goals to restore and sustain the health, productivity, and biological diversity of the oceans.

Bibliography

1. FAO (2006) State of world aquaculture: 2006. Food and Agriculture Organization of the United Nations, Rome, FAO Fisheries Technical Paper # 500
2. NOAA (2005) Fisheries of the United States – 2003. NOAA, Washington, http://www.st.nmfs.gov/st1/fus/fus03/index.html
3. Worm B, Barbier EB, Beaumont N, Duffy JE, Folke C, Halpern BS, Jackson JBC, Lotze HK, Micheli F, Palumbi SR, Sala E, Selkoe EK, Stachowicz JJ, Watson R (2006) Impacts of biodiversity loss on ocean ecosystem services. Science 314:787–790
4. Ryan J (2004) Farming the deep blue. Board Iascaigh Mhara Technical Report, pp 82
5. Shumway SE, Davis C, Downey R, Karney R, Kraeuter J, Parsons J, Rheault R, Wikfors G (Dec 2003) Shellfish aquaculture – in praise of sustainable economies and environments. World Aquacult 34(3):15–19
6. Buck BH, Krause G, Rosenthal H, Smetacek V (2003) Aquaculture and environmental regulations: the German situation within the North sea. In: Kirchner A (ed) International marine environmental law: institutions, implementation and innovation, vol 64. Kluwer Law International, The Hague, pp 211–229
7. Ward LG, Grizzle RE, Irish JD (2006) UNH OOA environmental monitoring program, 2005. CINEMar/open ocean aquaculture annual progress report for the period from 1/01/05 to 12/31/05. Final report for NOAA grant No. NA16RP1718, interim progress report for NOAA Grant No. NA04OAR4600155, Submitted 23, Jan 2006. http://ooa.unh.edu
8. Langan R (2007) Results of environmental monitoring at an experimental offshore farm in the Gulf of Maine: environmental conditions after seven years of multi-species farming. In: Lee CS, O'Bryen PJ (eds) Open ocean aquaculture – moving forward. Oceanic Institute, Waimanalo, pp 57–60
9. Langan R, Horton CF (Dec 2003) Design, operation and economics of submerged longline mussel culture in the open ocean. Bull Aquac Assoc Can 103-3:11–20
10. Buck BH, Thieltges DW, Walter U, Nehls G, Rosenthal H (2005) Inshore-offshore comparison of parasite infestation in *Mytilus edulis*: implications for open ocean aquaculture. J Appl Ichthyol 21(2):107–113
11. Jeffs AG, Holland RC, Hooker SH, Hayden BJ (1999) Overview and bibliography of research on the greenshell mussel, *Perna canaliculus*, from New Zealand waters. J Shellfish Res 18(2):347–360
12. Island Institute (Sept 1999) The maine guide to mussel raft culture. Island Institute, Rockland
13. Scott N, Tait M (1998) Mussel farming – an expanding industry in shetland North Atlantic fisheries college. Fisheries information note no. 1, October 1998
14. Bonardelli J (1996) Longline shellfish culture in exposed and drift-ice environments. In: Open Ocean Aquaculture: proceedings of an International Conference, Portland, 8–10 May 1996, Marie Polk editor. New Hampshire/Maine Sea Grant College Program Rpt. #UNHMP-CP-SG-96-9, pp 235–253

15. Michler-Cieluch T, Krause G (2008) Perceived concerns and possible management strategies for governing wind farm-mariculture integration. Mar Policy 32(6):1013–1022
16. Brenner M (2009) Site selection criteria and technical requirements for the offshore cultivation of blue mussels. Dissertation. Jacobs University Bremen, Bremen
17. SubSea Shellfish (Dec 2004) Biology and innovation primer project 2004. http://www.freepatentsonline.com/EP1476011.pdf
18. Stanley S (2005) development of a submersible raft for shellfish aquaculture. US Department of Commerce National Oceanic and Atmospheric Administration Small Business Innovation Research (SBIR). Abstracts of Awards for Fiscal Year 2005, pp 12
19. Buck BH (2007) Experimental trials on the feasibility of offshore seed production of the mussel *Mytilus edulis* in the German Bight: installation, technical requirements and environmental conditions. Helgol Mar Res 61(2):87–101
20. Holmyard J (2008) Potential for offshore mussel culture. Shellfish News, 25, Spring Summer 2008
21. Buck BH, Buchholz CM (2004) The offshore-ring: a new system design for the open ocean aquaculture of macroalgae. J Appl Phycol 16:355–368
22. SB Mariculture (2008) http://www.sbmariculture.com/
23. Brehmer P, Gerlotto F, Guillard J, Sanguinède F, Guénnegan Y, Buestel D (2003) New applications of hydroacoustic methods for monitoring shallow water aquatic ecosystems: the case of mussel culture grounds. Aquat Living Resour 16(2003): 333–338
24. Van Nieuwenhove K, Delbare D **(**2008) Innovative Offshore Mussel Farming in the Belgian North Sea. www.vliz.be/imisdocs/publications/132623.pdf
25. Jeffs AG (2003) Assessment of the potential for mussel aquaculture in Northland NIWA client report: AKL2003-057, NIWA Project: ENT03101. National Institute of Water and Atmospheric Research Ltd, Auckland
26. Thompson NW (1996) Trends in Australasian Open Water Aquaculture. In*:* Open ocean aquaculture: proceedings of an International Conference. May 8–10, 1996, Portland, ME. Marie Polk editor. New Hampshire/Maine Sea Grant College Program Rpt. #UNHMP-CP-SG-96-9: pp 223–234
27. Stevens C, Spigel R, Plew D, Fredricksson D (Mar/Apr 2005) A blueprint for better mussel farm design. NZ Aquac 04:8–9
28. Plew DR, Stevens CL, Spigel RH, Hartstein ND (2005) Hydrodynamic implications of large offshore mussel farms. IEEE J Ocean Eng 30(1):95–108
29. Lovewell MA (2008) The fisherman. Martha's Vineyard Gazette. 28 Aug 2008. http://www.mvgazette.com/article.php?18129
30. Maine DMR (2009) The blue mussel in maine. Maine Department of Marine Resources. http://www.maine.gov/dmr/rm/bluemussel.html
31. Lander T, Barrington K, Robinson S, Mac Donald B, Martin J (2004) Dynamics of the blue mussel as an extractive organism in an integrated multi-trophic aquaculture system. Bull Aquac Assoc Can 104-3:19–29
32. Langan R (2004) Balancing marine aquaculture inputs and extraction: combined culture of finfish and bivalve molluscs in the open ocean. Bull Fish Res Agency Jpn Suppl 1:51–58
33. Richards JB, Trevelyan GA (Dec 2001) Mussel culture. In: Leet WS, Dewees CM, Kleingbeil R, Larson EJ (eds) California's living marine resources: a status report. California Department of Fish and Game, Yountville
34. Buck BH, Krause G, Rosenthal H (2004) Extensive open ocean aquaculture development within wind farms in Germany: the prospect of offshore co-management and legal constraints. Ocean Coast Manage 47(3–4):95–122
35. Michler-Cieluch T, Krause G, Buck BH (2009) Reflections on integrating operation and maintenance activities of offshore wind farms and mariculture. Ocean Coast Manage 52(1):57–68
36. Hatcher A, Grant J, Schofield B (1994) Effects of suspended mussel culture (*Mytilus spp*.) on sedimentation, benthic respiration and sediment nutrient dynamics in a coastal bay. Mar Ecol Prog Ser 115:219–235
37. Fabia G, Manoukian S, Spagnoloa A (Apr 2009) Impact of an open-sea suspended mussel culture on macrobenthic community (Western Adriatic Sea). Aquaculture 289(1–2):54–63
38. Prins TC, Smaal AC, Dame RF (1997) A review of the feedbacks between bivalve grazing and ecosystem processes. Aquat Ecol 31(4):349–359
39. Grant J, Bacher C (2001) A numerical model of flow modification induced by suspended aquaculture in a Chinese Bay. Can J Fish Aquat Sci 58:1003–1011
40. Lloyd BD (2003) Potential effects of mussel farming on New Zealand's marine mammals and seabirds: a discussion paper. Department of Conservation, Wellington, Vii: 34 p
41. Paul W (1999) Reducing the Risk of Open Ocean Aquaculture Facilities to Protected Species. In: NOAA (Ed.): National Strategic Initiative Project, Summaries 1999, Aquaculture Information Center – DOC/NOAA. 1–11