Physics World Discovery

Space Weather

Mike Hapgood

Introduction: what is space weather?

There is a growing awareness across the world that our civilization needs to be more resilient to natural hazards, especially large and dangerous events that might happen only once a century. Such events lie beyond most personal and organizational experience, and only science can offer the breadth of insights needed to recognize, and make realistic preparations for, such events. One obvious historical example was the violent 1980 eruption of Mount St Helens in the US. The science showed clearly that this volcano had a history of occasional violent eruptions, but some people who lived peacefully alongside the volcano for decades lost their lives. This example shows how important it is to communicate our scientific understanding of natural hazards and to provide convincing arguments that reach out to people at risk, especially where that understanding goes beyond personal experience.

Space weather is an important modern example of a risk that needs good communication to wider audiences, especially when it can cause severe adverse impacts on human activities. As its name implies, ‘space weather’ originates in space, and can often be traced to physical processes on the Sun. Violent eruptions on the solar surface can generate huge clouds of magnetized plasma and/or high-energy particle radiation, which can then propagate across interplanetary space and envelop the Earth. Satellites and spacecraft must therefore be designed to have a high level of resilience to these events, but the effects of the erupted material can also penetrate the magnetic fields and atmosphere that surround our planet, reaching down and into the surface of the Earth. These effects can then disrupt, and sometimes damage, a wide range of everyday technological systems that are often critical to the smooth functioning of modern societies—now recognized as ‘critical infrastructures’.

One key challenge in communicating the risk from space weather is that it is not a direct risk to human activities here on Earth, but instead a risk to our modern dependence on advanced technologies, most obviously reliable continuous access to electricity. Other natural hazards such as strong winds, extremes of temperature, flooding and even objects falling from space, can be directly perceived by human senses and recognized as an obvious threat to life and property. But the only direct manifestation of space weather is the aurora in the night sky, and the natural human response is to see the beauty rather than the danger (figure 1). Only science allows us

doi:10.1088/978-0-7503-1372-8ch1

Figure 1. Images from two space-hazard events. Left: the trail left when a 12 000 ton meteor passed through the atmosphere over Chelyabinsk in Russia on 15 February 2013, generating a shock wave that damaged buildings and injured 1 500 people. Right: bright aurora over Oxfordshire in England on 13 March 1989, part of a global event that disrupted power systems across the world. This included tripping of UK National Grid transformers in Norfolk and Cornwall, alerting a generation of UK power engineers to space-weather risks. Chelyabinsk image by Alex Alishevskikh—Flickr: Meteor trace, CC BY-SA 2.0, https://commons.wikimedia.org/w/index.php?curid=24726667. Aurora image by the author.

to see the aurora as the result of solar eruptions reaching the Earth and driving a cycle of plasma physics processes within the magnetosphere.

But those plasma processes also drive phenomena that can disrupt technological systems. They can create strong time-varying electric currents though the upper atmosphere, producing marked variations in the magnetic field in the surface layers of Earth. These changing magnetic fields can then induce geoelectric fields that generate extra and unexpected electric currents through power grids, thereby disrupting the normal distribution of electricity. These plasma processes can also dissipate large amounts of heat in the upper atmosphere, particularly over polar regions, which profoundly changes the global circulation of the upper atmosphere and leads to worldwide changes in its density and composition. The end result can be major changes in the global morphology of the ionosphere—the ionized component of the upper atmosphere—which can affect the propagation of radio signals through that region.

The high-energy particle radiation reaching our atmosphere from the Sun can also impact the electronic devices that now pervade technological systems. We know that the most intense bursts of solar energetic particles can reach the Earth's surface, with more than 70 such events observed since 1942. These events can increase the natural radiation background and hence the risk of errors in electronic devices. The most obvious risk is in aviation, since at flight altitudes the atmosphere only provides 25% of the natural radiation shielding available on the ground, but also at risk are the control systems embedded in critical infrastructures such as electricity, gas, water, rail and road.

The risks from space weather have become more serious as we have become more heavily dependent on advanced technologies, and the seriousness of the risk has been reinforced by reports of adverse impact from major space-weather events, notably in 1967, 1972, 1982, 1989, 2000 and 2003. These events have helped to improve our understanding of space weather, and to establish a space-weather community that brings together experts from disciplines spanning science, engineering, economics and policy development. However, complacency remains a major challenge for this community to address, particularly as there have been no significant events for over a decade. This means that many organizations operating systems at risk from space weather have limited, sometimes declining, appreciation of those risks.

The aim of this book is to communicate our current scientific understanding of space weather, and how we critically need to improve our understanding to mitigate the risks it creates for our Earth-based technologies. It will look into the future of space-weather research, and what forthcoming technological developments may open up new risks that will drive future research.

The physics of space weather

Space weather is the consequence of a chain of physical processes that enable energy and momentum from the Sun to drive variations in a number of terrestrial environments. This chain is central to our understanding of space weather and is shown in highly simplified form in figure 2.

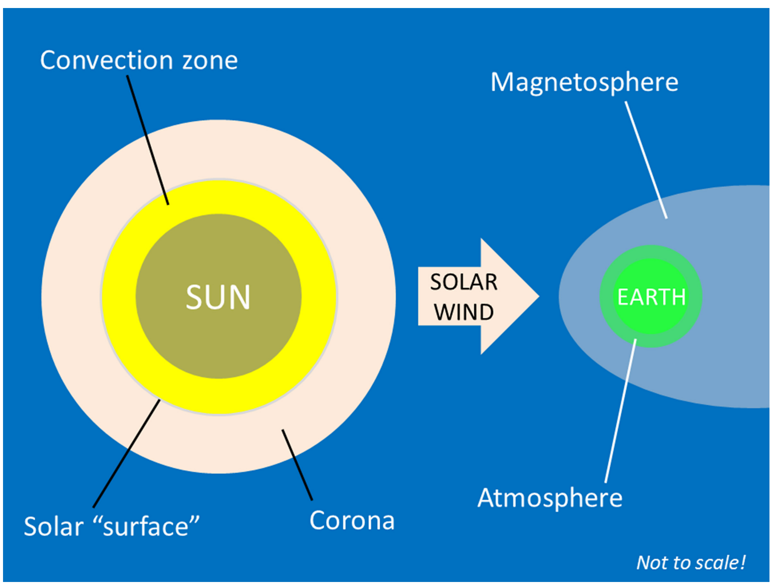

Figure 2. Schematic (not to scale) outlining the primary elements in the chain of physics through which the Sun drives space weather: the **convection zone** in the outer layers of the Sun (yellow) and the Sun's **corona** (pink); the outflow of plasma from the corona that we call the **solar wind** (arrow); Earth's **magnetosphere** (light blue), the region of space within which Earth's magnetic field that is confined by the flow of the solar wind, and the solid body (light green) and **atmosphere** (green) of the Earth. Schematic by the author.

The Sun: the engine of space weather

Space weather may be considered to arise in the outer layers of the Sun, where heat generated by nuclear fusion in the core is transported to the surface of the Sun by convection. The ionized hot matter in this convection zone is electrically conducting, so the convection flows generate magnetic fields through dynamo processes. As these magnetic fields rise to the surface of the Sun, they often become twisted into complex shapes—caused by the fact that different latitudes of the Sun rotate at different speeds (faster at the equator, slower at high latitudes). Magnetic field lines spanning a range of latitudes get stretched out and eventually form complex magnetic topologies containing significant free magnetic energy, and this is thought to be the main energy source for space weather.

The tangled magnetic fields rising out of the Sun confine dense hot plasma, as can be seen in extreme ultra-violet (EUV) images of the solar atmosphere, or the corona (figure 3). The release of energy from these magnetic field loops, which can happen through plasma processes such as magnetic reconnection (see box), drives a range of space weather phenomena. It can induce large-scale reconfigurations of the coronal magnetic fields, sometimes causing a cloud of plasma to be ejected into interplanetary space—a coronal mass ejection or CME. On smaller scales, the energy released by reconnection can create very high-energy electrons that can collide with the denser matter close to the Sun's surface to produce a burst of electromagnetic radiation, particularly at EUV and x-ray wavelengths, that's known as a solar flare. Thus CMEs and flares often occur in association, but are distinct phenomena that can also happen quite separately.

Both CMEs and solar flares contribute to the space weather effects that we see at Earth. However, flares dissipate their energy over a wide range of solid angles, so can affect us if they occur anywhere on the Earthward face of the Sun, while CMEs

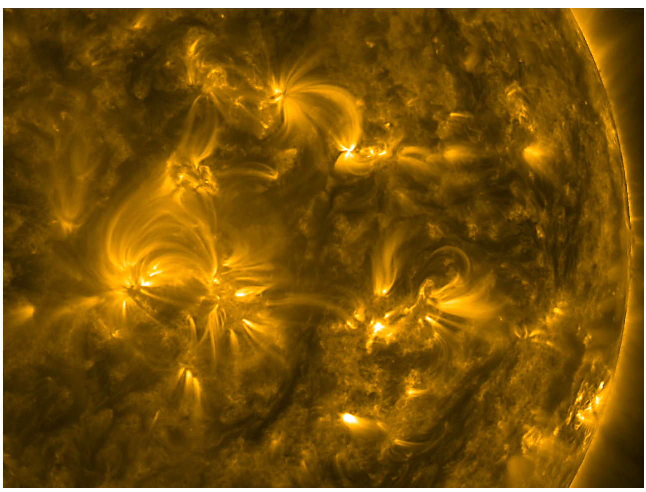

Figure 3. Extreme ultraviolet image of a tangle of arched magnetic field lines in the Sun's corona, taken in January 2016 by NASA Solar Dynamics Observatory. Credit: Solar Dynamics Observatory, NASA.

focus their energy into smaller range of solid angles. As a result, they often miss the Earth, but they provide a much stronger punch of energy when they hit our planet.

Plasma physics and space weather

An understanding of plasma physics is crucial to understanding space weather, with almost all of the regions shown in figure 2 filled with hot ionized matter. One key feature of these plasmas is that they are all electrically conducting—highly conducting in many cases as interparticle collisions are rare. The high electrical conductivity means that we can usually consider the magnetic field to move with the plasma, a concept known as 'frozen-in'. This frozen-in concept underpins many key ideas in the science of space weather, allowing us to visualize, for example, how the solar wind transports magnetic fields away from the Sun to create the spiral pattern shown in figure 5 later.

However, in some situations these frozen-in fields can break down, which can lead to an important plasma process called magnetic reconnection that can convert magnetic energy into kinetic energy. Reconnection can happen at the interface between two volumes of plasma with opposed magnetic fields, and which can have very different properties, such as at the edge of Earth's magnetosphere, where the very hot plasma of the magnetosphere meets denser cooler plasma from the solar wind.

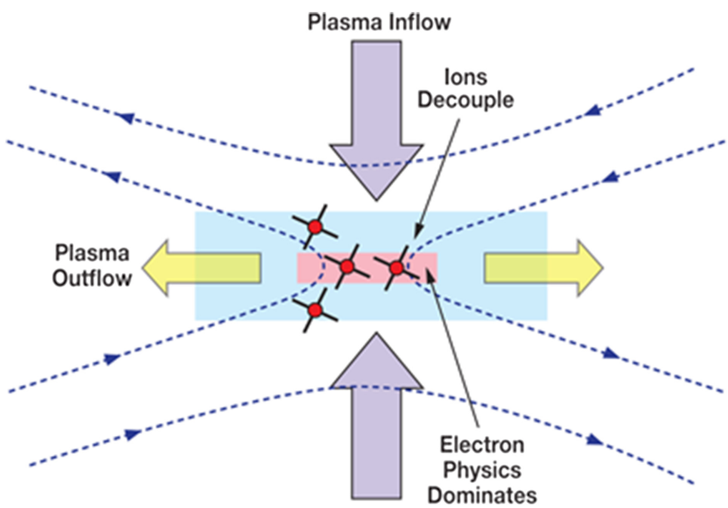

Figure 4. Schematic showing a 2D cut through a region in which magnetic reconnection is taking place between two plasma volumes at the top and bottom of the schematic. The magnetic field lines (dashed blue lines) have become interconnected between the two volumes, and plasma flows into the reconnection region from both volumes (purple arrows). Within the reconnection region, magnetic energy is converted to plasma kinetic energy in the form of outflow jets along the boundary between the two volumes (yellow arrows). A key scientific goal to help us understand the physics of reconnection is to make detailed plasma measurements within the heart of the reconnection region, as shown by the four satellites (red circles). Credit: Magnetospheric Multiscale, NASA.

The two fields are initially separated by a current sheet—as dictated by Maxwell's equations—but a breakdown of the frozen-in theorem can trigger a reconfiguration that causes the fields to become interconnected (see figure 4). In this new configuration, plasma and magnetic fields flow into the boundary from both sides, and then flow out along the boundary as shown by the purple and yellow arrows, with the outflowing plasma accelerated due to conversion of magnetic to kinetic energy. In the case of Earth's magnetosphere, these jets of outflowing plasma have been widely observed on a number of satellite missions over the past 40 years. The heart of the reconnection process—the tiny diffusion regions within which the frozen-in theorem breaks down and the magnetic fields are reconfigured—is at the centre of the X-pattern, and this region remains the object of intensive study as we strive to better understand reconnection.

The solar wind: the transmission system for space weather

The release of magnetic energy is also thought to be the mechanism that superheats the corona to around one million Kelvins, compared to 6000 K at the solar surface. Such high temperatures make the corona unstable against the gravitational pull of the Sun, allowing coronal material to escape the Sun to form the solar wind. In the equatorial regions of the Sun, the magnetic field structures usually act to slow this flow, which creates two main types of solar wind: (a) a low-speed solar wind from lower latitudes; and (b) a high-speed solar wind from polar regions, and also at lower latitudes when a 'coronal hole' forms that also allows outflows. When both low- and high-speed winds arise from regions at the same latitude but at different longitudes, the rotation of the Sun, combined with the radial flow of the solar wind, leads to a spiral pattern of solar wind streams in interplanetary space (figure 5). In some places the high-speed stream pushes into a low-speed stream, creating a region of compressed solar wind at the interface, known as a stream interaction region (SIR). This pattern of streams and SIRs is important for space weather because it modulates the solar wind impinging on the Earth, and also because it affects the propagation of CMEs away from the Sun.

When solar-wind plasma reaches the Earth, it is usually diverted by the Earth's own magnetic field, compressing that field on the dayside and stretching it out into a tail-like structure on the nightside, forming a diamagnetic cavity around the Earth called the magnetosphere (figure 6). However, this diversion will break down if magnetic reconnection occurs at the dayside boundary of the magnetosphere, which can happen if the orientation of the magnetic field embedded in the solar wind is opposite to that of the Earth's magnetic field. The two magnetic fields can then become linked, allowing solar wind plasma—and its associated energy and momentum—to enter the magnetosphere. This magnetic energy becomes stored in the tail of the magnetosphere, which then becomes unstable and creates another magnetic reconnection process that explosively releases the stored energy. Further energy inflows generate more magnetic energy in the tail, and in a major space-weather event this substorm cycle—also known as a Dungey cycle—will occur many times, with the ensemble of substorms being termed a geomagnetic storm.

Substorms: the Earth responds to space weather

Much of the energy released by substorms is transported Earthward and deposited in the atmosphere, where it drives a range of space-weather effects. Some of the

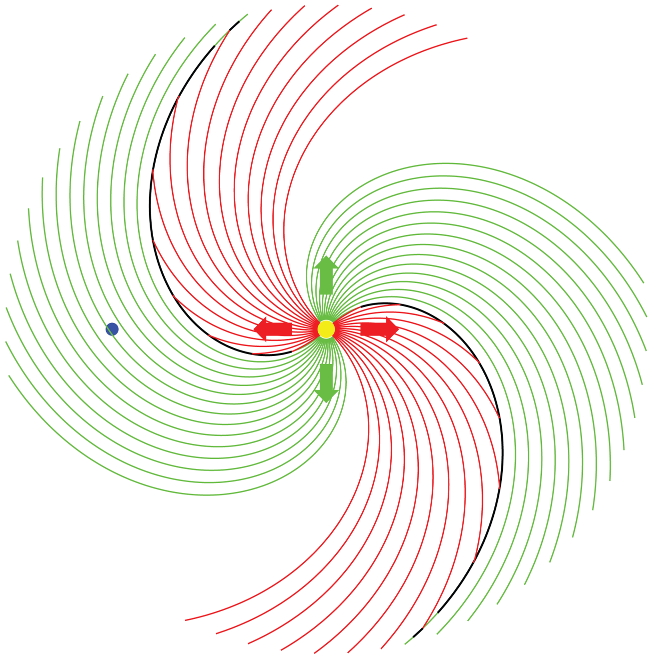

Figure 5. Schematic showing the spiral patterns that arise in interplanetary space when the Sun emits streams of fast (red arrow) and slow (green arrow) solar wind at different longitudes. Each parcel of solar wind flows radially away from the Sun, but its source region moves anti-clockwise with the rotation of the Sun, so that each stream and its embedded magnetic field (red and green lines) forms a spiral pattern. Where fast solar wind catches up with slow wind, a 'stream interaction region' of compressed solar wind is formed, as shown by the thick black line. If this pattern persists for more than one solar rotation, the SIR is also known as a co-rotating interaction region. Schematic by the author.

energy is transported by electric fields that propagate along magnetic field lines into the polar atmosphere, where they drive strong electric currents called auroral electrojets at altitudes between 100 and 140 km, a region where ions and electrons exhibit very different behaviours. As substorm energy is transported Earthward it also accelerates electrons to high energies, which also propagate along magnetic field lines to polar regions. Here they collide with the neutral atmosphere, mostly at altitudes of 90–150 km, producing optical emissions (the aurora), ionization and chemically active species such as nitric oxide.

Some substorm energy is dissipated in the magnetosphere, where it produces enhanced fluxes of both electrons and ions—particularly in the inner magnetosphere at altitudes between 12 000 and 25 000 km. The electrons and ions trapped in this region drift around the Earth in opposite directions (ions westward, electrons eastward) due to the curvature of the geomagnetic field, producing a ring of electric current around the equatorial regions.

Both the auroral electrojets and the ring current produce low-frequency variations in the geomagnetic field that penetrate deep into the solid Earth, generating geoelectric fields through magnetic induction. These fields produce electric currents

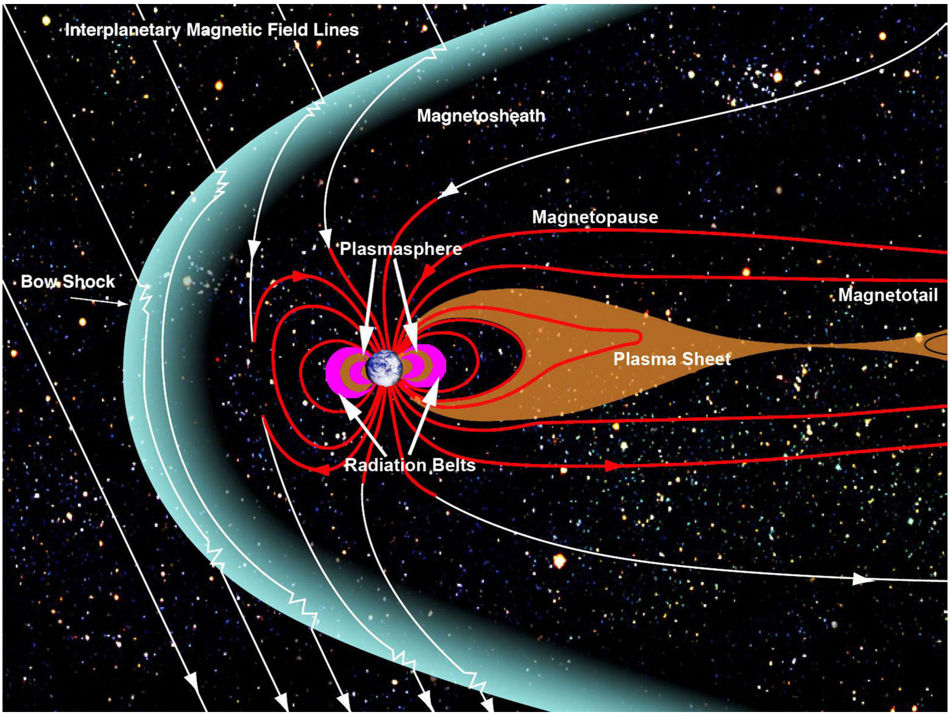

Figure 6. Schematic showing a 2D cut through the Earth's magnetosphere. The solar wind flows left to right and confines the geomagnetic field to a tadpole-like cavity, compressed on dayside and stretched out on nightside to form the magnetotail. A bow shock is formed upstream of the magnetosphere, so that the magnetosphere is encased in a region of shocked solar-wind plasma called the magnetosheath. In this case, reconnection is taking place on the dayside, so some magnetic field lines penetrate the magnetopause, the boundary between the magnetosphere and magnetosheath, and are shown as red inside the magnetosphere and white outside. Credit: NASA.

in the solid body of the Earth, and also in any man-made infrastructures that are electrically connected to Earth. Such Earth connections are central to the safe operation and electromagnetic compatibility of many extended infrastructures such as power grids and rail systems, but one side effect is that space weather can drive electric currents into those infrastructures. The low frequency of these geomagnetically-induced currents (GICs)—millihertz as compared with 50–60 Hz—means that they can drive grid transformers into half-phase saturation, causing effects such as voltage reductions and even physical damage. Since modern societies are critically dependent on the continuous supply of electricity, the effects of GICs on the power grid are generally considered by far the most important space-weather risk.

Energy from substorms is also a significant heat source for the polar upper atmosphere, and during strong geomagnetic storms it becomes the dominant heat source for the global upper atmosphere, overtaking the heat input from the Sun's ultraviolet rays (which is strongest at lower latitudes). This has a profound effect on the dynamics of the neutral component of the atmosphere, the thermosphere. Heat inputs together with Coriolis forces drive a global system of winds in the thermosphere, but during a major geomagnetic storm the global pattern of winds is transformed as the dominant heat inputs move from low to high latitudes. In addition, strong auroral heating can drive an upflow of molecular species such as nitric oxide and oxygen from below 100 km (where they are common) to much higher altitudes (where atomic oxygen is usually the dominant neutral species). The changes in these winds, and in the composition of the thermosphere, can in turn have major impacts on the ionosphere, as discussed below.

The ionosphere: at the heart of space weather

The interaction of the thermosphere with the ionosphere is a complex process, but the bottom line is that changes in the global circulation of the thermosphere have a profound impact on the global morphology of the ionosphere during large geomagnetic storms and for several days afterwards. The complexity arises because both 'spheres' occupy the same volume of space, and are weakly coupled via collisions between ions and neutral molecular species such as oxygen and nitric oxide. These collisions enable both momentum transfer from neutrals to ions—allowing the winds to drive plasma transport—and charge exchange, thereby creating molecular ions that will greatly increase recombination rates and lead to a loss of ionization (see box). Such an increase in recombination rates sometimes causes the near-total disappearance of the night-time ionosphere at mid-latitudes.

As with any plasma, the electrons in the ionosphere have a natural frequency at which they oscillate with respect to the much more massive ions. Electromagnetic waves with frequencies below the plasma frequency are damped and cannot propagate through the plasma, and will be reflected if they enter the plasma from an external source. Waves with frequencies above the plasma frequency do propagate, but more slowly than usual.

In the ionosphere the plasma frequency varies between 1 to 20 MHz, depending on a range of factors including time of day, season of year and phase of the solar cycle, as well as latitude and longitude. Radio waves at megahertz and lower frequencies are therefore reflected from the ionosphere, which has enabled various technologies for communications and surveillance that send radio waves over the horizon. Choosing the optimal frequency for reflection requires knowledge of the current plasma frequency around the reflection point, and in quiet space-weather conditions these optimal frequencies can be forecast from the regular patterns at any particular location. But a large geomagnetic storm disturbs these regular patterns—particularly during the night, when the ionosphere can then disappear completely—and so disrupts these technologies and can even cause them to fail completely.

For satellite technologies, radio transmissions to and from the spacecraft must operate at higher frequencies to enable them to propagate through the ionosphere, ranging from VHF systems at around 100–200 MHz up to Ka band (26–40 GHz) for high-capacity communications links. The ionospheric delay can be a major issue for some of these radio technologies, since they need to measure the distance between a spacecraft and a receiver. The obvious example is a GPS receiver, which determines its position by accurately measuring the distance from the receiver to several GPS satellites. Ionospheric delay is the main source of uncertainty in GPS position measurements, capable of introducing errors of many tens of metres, and so ionospheric models and measurements are widely used to correct for these errors, particularly during geomagnetic storms.

Another important feature of the ionosphere is that, like any plasma, it is prone to a wide range of small-scale instabilities. These occur particularly in the polar and equatorial regions and become more significant during geomagnetic storms. The instabilities cause refractive index variations that act like a diffraction screen, so that radio signals passing through these instabilities are subject to phase and amplitude scintillation. Ionospheric scintillation particularly affects frequencies below 3 GHz, where it can disrupt signal reception and thereby interrupt services that rely on radio signals. For example, the location and timing services provided by GPS are liable to interruption whenever strong ionospheric scintillation occurs, which means that GPS applications must be able to cope with such disruption.

Recombination in the ionosphere

Loss processes are a key element in the dynamics of the ionosphere. Ionization is usually lost by recombination, which in this case is the interaction of an ion with an electron to produce one or more neutral species. In the Earth's ionosphere we are particularly concerned by the recombination of oxygen ions (O^+), which is the dominant ion species at altitudes of 150–400 km. These ions can be combined with electrons to create neutral oxygen atoms via two very different routes:

Radiative recombination that involves direct loss of an O^+ ion: $O^+ + e^- \Rightarrow O + \gamma$, where γ is a photon.

Dissociative recombination, a two-step reaction in which an oxygen ion is initially converted to NO^+ by charge exchange with neutral NO molecules, $NO + O^+ \Rightarrow NO^+ + O$, then this ion recombines to produce separate N and O atoms: $NO^+ + e^- \Rightarrow N + O$.

Dissociative recombination has a much higher cross section because each step produces two particles, which means that excess energy from the reaction can be converted into the kinetic energy of the particles, while conserving momentum. Thus recombination rates in the ionosphere are increased when substorm activity drives molecular species such as NO and O_2 to higher altitudes.

Solar flares: fascinating science, limited impact

Solar flares—the spectacular bursts of electromagnetic radiation from the Sun's surface—can produce a number of space-weather effects across the dayside of the Earth as radiation from the flare reaches the Earth's atmosphere. Extreme ultraviolet (EUV) radiation from the flare produces additional ionization in the ionosphere at heights above 150 km and that will persist for many hours, thus changing the ionospheric delay of GPS signals. Modern GPS receivers can correct for this extra ionization within ten seconds, but for a brief period at the onset of large flares the receiver may lose track of the GPS signal delay. These ‘cycle slips’ are regular features of GPS systems that arise from a number of causes, which means that GPS receivers are designed to recover quickly from them—with the result that flares have a minimal impact on GPS usage.

The radiation reaching the Earth from solar flares also includes x-rays, which penetrate deeper into atmosphere to produce additional ionization below 100 km. At these altitudes ionospheric absorption becomes an issue, particularly when the frequency of collisions between electrons and neutral atmospheric species is higher than the frequency of a radio wave passing through that part of the ionosphere. When the radio wave excites oscillations of free electrons, those electrons lose energy via collisions rather than re-radiating the wave energy, which causes the radio wave energy to be absorbed. This particularly affects radio waves at frequencies of 1–20 MHz, as used for over-the-horizon communications and surveillance, and the absorption caused by a large solar flare can disrupt these systems and in the worst cases cause a blackout for an hour or two across the dayside of the Earth. Although this was a major impact of space weather in the mid-20th century, when these systems were a mainstay of long-distance communications, it has become less important with the introduction of satellite communications over the past 40 years.

However, solar flare x-rays still have a significant impact on civil aviation, where the ability to use megahertz frequencies for over-the-horizon communications is valuable when flying over remote regions, and is mandatory on some routes, especially over the oceans. However, a range of satellite-based communication services are now providing an alternative for aviation, and have initially been introduced on the busiest routes that are home to the most modern aircraft. As these satellite services continue to expand and evolve over the next decade or two, the space-weather impact of solar flare x-rays is likely to reduce to just a nuisance level.

Radiation: the unseen hazard from space

The space around the Earth is full of highly energetic particles, including a slowly changing background of galactic cosmic rays. But, from time to time, intense bursts of particle radiation generated by plasma processes in the solar corona and in interplanetary space can greatly increase this background radiation. One important example is the shocks that form in the plasma ahead of fast CMEs as they leave the Sun. These shocks can generate large fluxes of energetic ions in the energy range from 10 MeV to 10 GeV, and will sustain those fluxes for several days as they cross interplanetary space, creating a long-lived radiation storm. Solar flares can also be accompanied by short bursts of intense radiation, lasting for anything from a few hours up to a day. In this case the radiation is thought to be generated by processes associated with the flares, perhaps local shocks linked to magnetic reconnection, and the radiation will reach interplanetary space only if the particles can reach magnetic field lines that are open to interplanetary space.

Once energetic charged particles reach interplanetary space they propagate outwards following the interplanetary magnetic field, which typically has the spiral structure shown in figure 5. Thus those particles that reach Earth will typically approach from a direction 45° west of the direction of the Sun, and with some spread around that direction due to the natural gyration of charged particles around magnetic field lines. However, as a radiation event proceeds, irregularities in the interplanetary magnetic field cause the charged particles to scatter. Thus the angular distribution of the particle radiation quickly becomes isotropic, most importantly including the backscatter of particles towards the Sun. This means that there is no way to hide from these particles, and satellites must therefore be designed to be resilient to these radiation events, within the limits of affordability. For example, solar arrays and other components might be designed to survive radiation damage with a 95% probability that the allowed radiation dose will not be exceeded during the planned satellite lifetime (figure 7).

Energetic particles can also cause single-event effects by depositing charge inside a digital device, upsetting or disrupting its operation. Such events might just cause bit-flips that corrupt data and software, but more complex effects have emerged as digital devices have become more sophisticated, ranging from latch-up—where devices become stuck in a particular state—to effects that cause physical damage to the device. Some of these effects can be mitigated by engineering methods, such as the use of error correction codes to detect and fix bit-flips, but many require operator

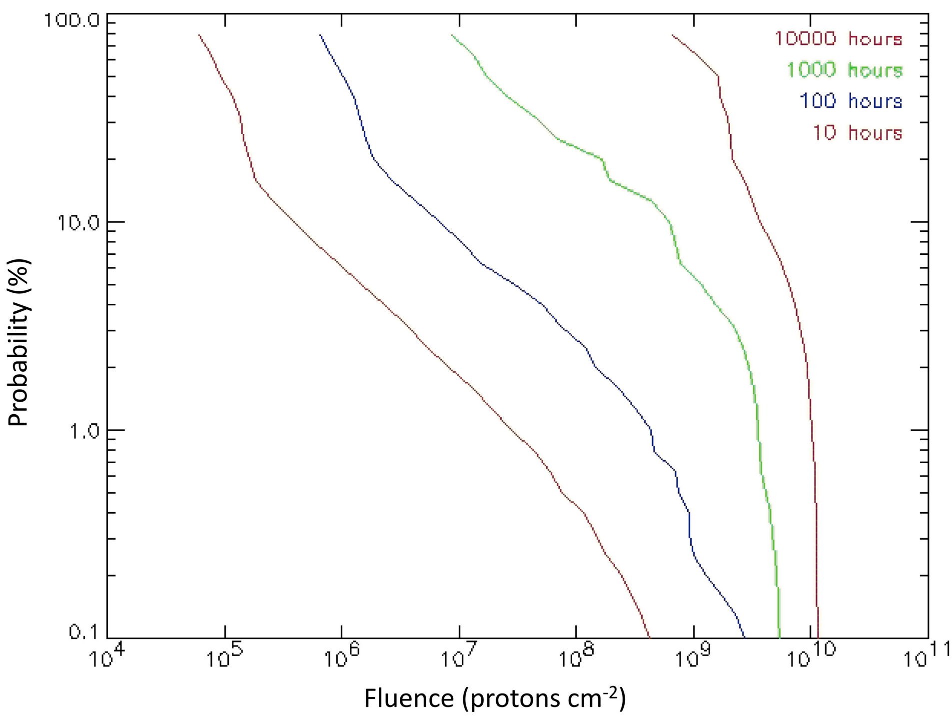

Figure 7. Plot showing levels of >10 MeV proton fluence (i.e. time integral of proton flux) likely to be accumulated on a satellite near Earth, but outside the protective shield of the geomagnetic field. The four coloured lines show the probability of the fluence reaching a particular level, for each of four time periods from 10 h to 10 000 h. Plot by the author.

action to reset devices and recover normal operations. Dealing with such single-event effects can therefore be a significant part of the work of satellite operations teams, with serious workload implications during a major space weather event.

Large radiation storms can also have a major impact on aviation, particularly for routes over the Arctic. This is because particle radiation reaching the Earth interacts with the geomagnetic field, which has the effect of focusing particles with energies lower than about 1 GeV towards the poles. These particles can create ionization at altitudes below 100 km, which, like the x-rays from solar flares, can increase ionospheric absorption. The result can be a blackout of megahertz radio communications in polar regions, which is crucial for the safe operation of Arctic routes as there is currently no reliable satellite communications service in the high Arctic. This blackout in the polar regions can last several days during a long-lived radiation storm, and for its duration the busy Arctic air routes between the USA and major Asian countries such China and India must be closed.

Radiation storms can also include much higher energy radiation that can reach lower latitudes and lower altitudes, even right down the Earth's surface—as observed on more than 70 occasions since the first ground-level event was detected in 1942. The effects of this atmospheric radiation are a particular concern at aircraft flight altitudes of 10–12 km, since a large radiation storm can increase the background radiation by a factor of 100. This can cause single-event effects in digital devices used in critical aircraft systems such as engine and flight control, which makes it vitally important to design and test aircraft systems to ensure they are resilient to radiation. There is a longer term human risk too, since the extra radiation exposure could lead to small changes in the long-term cancer risk for the crew and passengers.

The role of research in mitigating space-weather risks

Space weather is a natural hazard, so we can and should use scientific knowledge to understand how it affects human societies and to identify the actions we can take to reduce the risks. Thus we should not treat space-weather research solely as an academic exercise, since there is also much value in considering, from the outset, how research results may help to mitigate the adverse impacts of space weather. To do that it is vital that researchers have some understanding of the routes through which space weather may be mitigated.

The most important is good engineering, since the major adverse societal impacts of space weather occur when critical infrastructures are disrupted. Operators of those infrastructures can respond by designing out the vulnerability to make the infrastructure more resilient to space weather, just as they do for other risks such as normal weather. When done well this approach can be very effective, as shown by the example of satellite navigation systems that within about ten seconds can correct for most of the position errors that arise from the effects of solar flares.

This makes it important to understand how research can help the design of engineering resilience. Here the critical need is to establish the worst-case environments that should be considered by engineers, not just today but also into the future.

The timescales for such future-proofing depends on the system being built; in some cases it might be just a few years to cover the planned operational lifetime of the system, but longer timescales may be chosen for other reasons. As an example, the insurance industry frequently considers risks at the 1-in-200 year level, whilst some critical systems are legally required to consider risks at the 1-in-10 000 year level. Only science can provide robust worst-case scenarios over all these timescales by exploiting historical and proxy data together with physics-based simulations. In contrast, personal and organizational experience is generally limited to the past one or two decades.

Given these scenarios, engineers can design systems with a low probability (typically a few percent, but ultimately set by affordability) that space-weather conditions will exceed design limits. Such systems should cope with the majority of space weather effects, but there will always be cases where design limits are exceeded. In these cases, a second route to mitigation becomes very important, which is the ability to forecast future space-weather conditions. Forecasting is particularly valuable when operators can configure their systems to temporarily improve resilience. Power grids, for example, can benefit from an ‘all-on’ policy, where routine maintenance is cancelled to ensure high availability across the generation and transmission systems, so that the grid has the best possible redundancy to cope with any problems.

However, it is important to recognize that operators of vulnerable systems ultimately need a forecast of the environment impacting their system. To be effective, forecasts of any element in the chain of physics linking the Sun to the Earth should therefore address what it means for the environment of interest to the operator. For example, if a CME is observed leaving the Sun, can we forecast the resulting geoelectric field and how that field will affect the power grid?

Such end-to-end forecasting requires a set of models that link different parts of the chain. Ideally these would all be physics-based, but that is not yet fully feasible today and some of the links may be based on empirical models. In principle, these sets of models may be run in a deterministic manner, but in practice a statistical approach is a better long-term aim, in line with modern forecasting of other environments such as the Earth’s weather. Here, ensemble methods are now widely used, essentially running a model many times with small changes in the input parameters that reflect the uncertainty in those parameters (including empirical parameters within models, as well as parameters that reflect current conditions). By obtaining an ensemble of model outputs and the distribution of values in those outputs, it is possible to estimate the statistical likelihood of particular outcomes, given the uncertainties in the input parameters. Human forecasters can then apply their experience to make predictions that indicate a range of possible environment conditions, including some estimate of the likelihood of adverse conditions.

Another valuable tool in the mitigation of space weather is nowcasting—in other words, providing information on current space-weather conditions, rather than forecasting what may happen in the future. Nowcasting is valuable when operators have limited ability to configure their systems to improve resilience, but need to recover quickly after a problem has arisen. It allows operators to quickly check whether the problem could have been caused by space weather, allowing them either to take appropriate actions for a space-weather event or to explore other causes.

Finally, it is important to recognize that governments are crucial players in the mitigation of severe space weather—just as they are for other severe hazards. Since many of the systems sensitive to space weather are critical infrastructures, a severe event can trigger a cascade of effects that impact wider society and the economy. Good emergency management by government, not least coordinating efforts of both public and private sectors, can reduce both the scale of these secondary effects and the time needed to recover from them. For an effective response, emergency managers in government need to have good situational awareness when severe space weather occurs, ensuring they are aware of the threat before it happens and of how the space-weather conditions are likely to evolve in the lead-up to, and during, the event. These managers require all of the tools available—including worst-case scenarios, forecasts and nowcasts—and therefore governments worldwide are a major customer for the outputs from space-weather research.

Some critical issues in space weather research

The previous sections have provided an overview of our scientific understanding of space weather and the need for mitigation. We now build on these by outlining four examples of where better understanding is critically needed in order to deliver better mitigation of space-weather risks. These examples will trace backwards from space-weather impacts on Earth to the energy sources of space weather on the Sun, since this helps us to focus on the chains of phenomena that cause major space-weather events—and hence deliver new results that are useful, as well as exciting. It reduces the risk that we get diverted by phenomena that are fascinating, but that have only minor space-weather impacts.

Power grids: understanding substorms and ground conductivity

Electrical power is the fundamental infrastructure of modern societies and, for this reason, governments around the world are alert to anything that threatens its resilience. Space weather very clearly falls into this category, which requires good forecasts of when and where strong geoelectric fields, and hence strong GICs, will arise.

However, our ability today to make regional forecasts of geomagnetic variations, let alone geoelectric fields or GICs, is very limited. We rely mainly on a variety of systems offering global warnings of geomagnetic activity or GICs, largely based on numerical methods such as neural networks. It is widely recognized that advances beyond this state-of-art will require operational use of physics-based models, opening a path through which we can raise forecast skills by iteratively improving the representation of physics in these models—just has already been done in meteorology.

A first step towards this has recently been achieved in the US, where the Space Weather Prediction Center in Boulder, Colorado, has introduced an operational model to predict local geomagnetic variations using a magneto-hydrodynamic

(MHD) model of the Earth's magnetosphere. This major development follows extensive work to evaluate existing physics-based models of the magnetosphere, but there is considerable scope to further improve how physics is included in the simulations. Key challenges include a more realistic representation of the ionosphere and its coupling to the magnetosphere, and of the plasma as a mixture of charged particles rather than a conducting fluid. The latter is a fundamental step, as in any other plasma physics modelling, since it allows critical processes such as reconnection to be included in a self-consistent way. These new approaches will require more modern methods for numerical modelling, including adaptive methods in which the grid size, and even the depth of physics representation, is varied across the region being modelled. This enables the simulations to focus in detail on regions of particular physical significance, such as the tiny regions where reconnection occurs, while dealing efficiently with large regions where conditions vary more slowly.

Improved modelling is also needed to calculate the geoelectric fields induced at the surface of the Earth for a given set of geomagnetic variations. There are a number of well-established approaches to such calculations, which model how electromagnetic variations propagate into the conducting body of the Earth. Here a key challenge is to take account of small-scale variations in the conductivity of the Earth, which arise both from the complexity in the subsurface geology and from the abrupt change in conductivity between land and sea. Work to address this challenge will likely require deeper application of numerical methods to enable us to model the effects of induction in the complex, sometimes fractal, geometries that characterize geology and coastlines. The latter is particularly important because many countries have much of their critical infrastructure in coastal regions.

Atmospheric radiation: a concern for aviation

Only around 8% of the natural background radiation on the surface of the Earth originates from space, in the form of cosmic rays, but that contribution increases with altitude. At aircraft cruise altitudes, cosmic rays completely dominate the radiation environment, producing a natural background that is 30–40 times greater than on the surface. As a result, aircraft crew are now recognized as having the greatest occupational exposure to radiation, and so their employers are required to assess their cumulative radiation exposure and take action if it exceeds regulatory limits.

Their exposure is usually assessed by modelling the radiation levels for the flights on which any individual has worked, since the monitoring of individual staff is considered impractical in the aviation environment, and the deployment of systems for on-board monitoring of individual aircraft is still in its infancy. Thus there is a need for accurate models of the atmospheric radiation environment with sufficient spatial and temporal resolution to allow reliable calculations of radiation fluxes along the flight track of any aircraft.

These models must integrate a number of factors, including:

a. the collisions of energetic particles with the molecules that form the bulk of the neutral atmosphere, the consequent production of secondary particles (mainly neutrons and muons), and the vertical and lateral transport of these particles through the atmosphere;

b. the deflection of energetic particles approaching the Earth by electromagnetic forces as they pass through the geomagnetic field before reaching the atmosphere;
c. the changes in this deflection during major geomagnetic storms, since the enhanced electric currents that characterize these storms significantly change the geomagnetic field;
d. changes in the flux of energetic particles reaching the Earth, including both the reduction of these fluxes as CMEs pass over the Earth (a shielding effect known as a Forbush decrease) and the huge flux increases during solar radiation storms.

There are already a number of models that address these issues and, most importantly, growing efforts to validate those models against actual measurements. However, these measurements are still relatively sparse, as there is very limited deployment of radiation monitors on aircraft. But their importance for advancing the modelling of atmospheric radiation has prompted strong research interest in developing monitoring equipment, and in working with the aviation industry and regulatory authorities to expand measurement programmes. Particularly lacking are measurements made when atmospheric radiation is enhanced by solar radiation storms, so researchers are developing methods to rapidly deploy additional radiation monitors during an event, such as by using balloons.

For aircraft passengers, the cumulative exposure to background radiation is much lower because they generally spend less time in the air, and so the risk from radiation is usually much smaller. However, it is possible to accumulate significant radiation exposure on a single flight, particularly during an intense solar radiation storm containing significant fluxes of particles with energies as high as a few GeV. In this case, the exposure would need to be assessed using an atmospheric radiation model, ideally complemented by some simultaneous measurements to validate the model outputs, to enable proper advice to be given to the passengers. This reinforces the need to develop and sustain measurements of atmospheric radiation, as well as to develop models of that radiation.

Hot plasmas and killer electrons

All objects in space, both natural and man-made, are electrically charged. They are constantly exposed to the electrons and ions that form the plasmas that fill all space environments. The lighter electrons penetrate more deeply into surface materials, leading to an accumulation of negative charge on the surface of the object. When the object is in sunlight, this is countered by photoemission of electrons, leading to an accumulation of positive charge. The net charge on the surface is determined by the interplay of these two factors. If the flux of UV photons from the Sun dominates over the flux of electrons from the local plasma environment, that net charge will be positive. But if the flux of electrons dominates, the net charge will be negative. This most obviously happens when the object is eclipsed by another body, preventing any photons from the Sun from reaching the object.

The flow of electrons into and out of the surface of a space object constitutes an electric current flowing between the object and its local plasma environment, causing the object to have an electrical potential relative to that environment (which may be regarded as the local zero potential, akin to the way that the potential of the ground provides the local zero potential for electrical systems on Earth). The electrical potential of the object will tend towards an equilibrium, such that the current flow between the object and local plasma is zero. When photoemission dominates, this equilibrium potential is a few volts positive, as most photoelectrons have energies of a few electron volts. But when the flux of plasma electrons dominates, the equilibrium potential will tend to a negative voltage determined by the mean energy of the plasma electrons. In many space environments, such as the geosynchronous orbit used by many operational satellites, the mean energy of the plasma electrons can be extremely high, which means that objects in these environments can acquire electrical potentials of hundreds or thousands of volts.

Such high voltages are potentially dangerous if they lead to differential voltages between parts of the spacecraft, thereby creating conditions in which electrical discharges can occur. These discharges can create false signals that disrupt satellite operations, and in the worst cases cause damage to satellite systems. Satellite manufacturers are well aware of this threat, and so take great care to ensure that all surfaces at risk are electrically bonded to the satellite chassis. This bonding does not require high conductivity—just enough to prevent charge accumulating to produce 'hot spots' where discharges could occur. The key point is to avoid exposing highly dielectric materials to the hot plasmas that are often found in space. If their use is essential, they should be protected, such as with a conductive layer that can drain electric charge.

However, some space plasma environments are so extreme that the charging problem is not limited to surfaces. Electrons at energies of millions of electron volts can penetrate deep into satellite systems and deposit charge inside any dielectric material they encounter, such as circuit boards and insulation on electric cables. If a satellite is exposed to these MeV electrons over a long period, the charge inside the dielectric can accumulate and may not escape until the resulting electric field causes a breakdown of the dielectric—leading to repeated electric discharges inside the material. This process is sometimes prosaically termed 'internal charging', but 'deep dielectric charging' is a much better description of the physics at play, and it is beautifully shown in a benchtop demonstration that can be viewed on-line at https://www.youtube.com/watch?v=eCz7BL74D4Y.

As with surface charging, these discharges can generate false signals that cause satellites to misbehave, and in the worst case they can damage satellite systems. Unfortunately MeV electrons are a common occurrence at the geosynchronous orbit used by many communications and meteorological satellites, and they are even more common on the so-called 'middle Earth orbits' used by navigational satellites such as the GPS and Galileo constellations. What's more, deep dielectric charging inside a satellite is hard to mitigate by engineering design. There is no natural balance due to emission of photoelectrons when in sunlight, and it is difficult to create low conductivity paths that can dissipate charge from deep inside dielectrics.

As a result, satellite operators need to be ready to deal with anomalous behaviour when there are substantial fluxes of MeV electrons in geosynchronous and middle Earth orbits, and increasingly they make good use of forecasts of MeV electron fluxes. They are also very supportive of research to improve these forecasts through the introduction of physics-based modelling. Many researchers are now developing models of electron energization and transport inside the Earth's magnetosphere, and how these are affected by the strength of the solar wind flowing past the Earth.

Both surface charging and deep dielectric charging can lead to discharges that disrupt or damage satellite systems, but it is important to recognize that they are caused by different processes:

a. surface charging is due to the interaction of the satellite with the hot plasmas found in many space environments;
b. internal or deep dielectric charging due to very high-energy electrons that deposit charge deep inside dielectric materials within a satellite

These different causative processes drive different mitigation paths. Surface charging can be reduced by careful engineering design and construction, ensuring that all satellite surfaces are conductive and electrically bonded to the satellite chassis. In contrast, deep dielectric charging is hard to mitigate by design, so satellite operators must be ready to take corrective action to counter anomalous satellite behaviour—which in turn requires good use of forecasts that predict periods of elevated high-energy electron fluxes.

Discussions about space-weather impacts on satellites sometimes fail to distinguish the forms of charging, instead just considering charging as a general issue. This is a dangerous misunderstanding as it can lead to inappropriate mitigation of the risks from space weather. It offers an excellent example of why it is vital to understand the physics of space weather.

Great balls of fire: how the worst space weather reaches Earth

The majority of adverse space weather arises from the interaction of the solar wind with the Earth's magnetosphere, most notably when magnetic reconnection enables solar-wind energy to drive geomagnetic activity, but also via other processes that drive instabilities and waves in the plasma and the magnetic fields that fill the magnetosphere. Thus we need to understand how the solar wind transports energy from the Sun to the Earth, and in particular to forecast the solar-wind conditions impacting the Earth. Such forecasts need to be coupled to models of the near-Earth environment to predict the conditions that can affect vulnerable technologies such as power grids and satellite navigation, but the solar-wind forecasts are an absolute pre-requisite for these wider predictions.

Thus solar-wind forecasting is an area of active research and one with many elements. The most crucial elements are to forecast (a) the speed and density of the solar wind, and (b) the strength and orientation of the magnetic field embedded in the solar wind. The speed and density are vital because it is the bulk motion of the solar wind that transports energy from the Sun to the Earth (only a few percent of solar wind energy is contained in the magnetic field). But the magnetic field, and especially its orientation, is nonetheless vital as it determines how this energy is coupled into the magnetosphere to drive geomagnetic activity.

The central focus of solar-wind forecasting is the propagation of CMEs through interplanetary space, since the highest solar wind speeds and densities (and indeed magnetic fields) are usually associated with CMEs. The simplest models assume that CMEs follow a ballistic trajectory once they have left the vicinity of the Sun. But interplanetary space is not a true vacuum, it is full of other solar-wind plasma that can interfere with the propagation of CMEs. If the CME is faster than the existing solar wind, for example, it will sweep up plasma from the background, forming a sheath of compressed plasma in front of it, generating a drag force that slows down the CME. And if the CME overtakes some existing structure in the solar wind, the resulting interaction can sometimes produce enhanced densities and magnetic fields, leading to stronger geomagnetic activity at the Earth. A recent example is the St Patrick's Day geomagnetic storm of 2015, where the interaction between a CME and an SIR is thought to have caused a higher level of geomagnetic activity than was expected. Though not a large storm by historical standards, it was the largest event for several years and was well-observed with modern instruments. It showed us that the evolution of CMEs during their journey to Earth is an important factor in driving major space-weather impacts at Earth.

Today we have a mix of parametric and physics-based models of CME propagation through the solar wind. Parametric models have proved a useful forecasting tool for straightforward cases, such as a single CME, and today they form the basis of operational services to predict CME arrival times. These models typically describe the CME by a simple geometric shape, most commonly similar to an ice-cream cone (that is, a spherical cap on top of a cone). In this case, the cap surface represents the leading edge of the CME, while the cone represents the magnetic fields stretched out behind it that link the CME back to the Sun. A range of cone models handle the expansion of the CME as it moves away from the Sun in different ways; key questions include whether the spherical cap maintains the same shape (self-similar expansion) or a constant radius, and whether the cap propagates at constant speed or is slowed by drag.

In contrast, physics-based models numerically simulate the behaviour of the solar wind over the whole of the inner solar system out to beyond the orbit of the Earth, and even further when it is desirable to model CME propagation past the outer planets—for example, to compare with observations on deep-space missions. These models must be initialized with an assessment of the existing solar-wind conditions across the grid used in the simulation, and an assessment of the CME launched from the Sun. The model can then be run to simulate how the CME will evolve as it propagates away from the Sun. The outcome of these simulations depends on a number of factors, in particular the amount of physics included in the model and the quality of the initial conditions. Current models are based on fairly simple approaches—for example, an ideal MHD description of the plasma in the solar wind—but they are proving very useful as a tool for front-line operational forecasting centres. They can give forecasters not just arrival times, but also a broader overview of the state of the solar wind and hence useful context for the human assessment of space-weather conditions.

But there is considerable scope for improvements to these physics-based models—most obviously by including deeper description of the plasma physics—which could be important for modelling the very fast CMEs that cause extreme space-weather events. Measurements suggest that suprathermal particles contribute significantly to the plasma pressure within and around such CMEs, which means that we need to develop models that can incorporate this and other plasma effects. Another important area for improvement is better initialization, both of the CME and the existing solar wind, and a move to assimilative modelling, whereby observations of CMEs in flight to Earth can be used to update forecasts. These are both active areas of research that link observers and modellers, with the aim of exploiting the wealth of observational techniques now available to improve initialization and/or assimilation, and, most important, to test the outcomes against real CMEs.

Improved physics-based modelling opens up further important possibilities in forecasting the solar wind. Most significant is the possibility to forecast reliably the orientation of the interplanetary magnetic field—which would be a profound step forward in our ability to predict whether a particular CME will lead to a strong geomagnetic storm. It is not possible with current operational models, but is a major focus for current research—and one that could bring major rewards. Almost as important is the ability to model interactions of CMEs with other CMEs or SIRs, and hence whether that interaction could lead to a stronger than expected geomagnetic storm. Clearly this is beyond the capabilities of parametric models, but physics-based models have the potential to offer new insights as long as the grid resolution is sufficient to model the interaction on a small enough scale.

Finally, we consider the production of shocks by CMEs, which can happen if the CME is moving sufficiently fast in the background solar wind. The shock forms in front of the CME, and the ambient solar-wind plasma swept up ahead of the shock will form a sheath of heated plasma between the shock and the CME. As a result, these shocks are major sources of high-energy charged particles, and hence of solar radiation storms. Modelling the propagation of CMEs may therefore help us to model radiation storms, which makes this another important research objective.

Looking to the future

Today, we are at an exciting stage in the development of space-weather services. Traditionally these services have been based on empirical models such as neural networks, but we are now in the first stage of a transition to physics-based models. A number of these new models are now in operational use, notably the WSA-Enlil solar-wind model used by the SWPC to predict the arrival at Earth of solar-wind disturbances such as CMEs. But this is just the first phase of a long-term transition in which the accuracy and usefulness of space-weather services will be enhanced through better physics. That development requires a creative tension between improved models of the physics, and validation of the models against observations (such as that promoted in the US through the Coordinated Community Modeling Center, a multi-agency initiative hosted at NASA's Goddard Flight Center).

Thus there are considerable opportunities for space-weather research: in understanding the physics of space weather, in improving how that physics is included in models, in making more extensive measurements, and in comparing models and measurements. For example, in the short- and medium-term we need to improve models of how CMEs and SIRs propagate to Earth and the consequent response of Earth's magnetosphere and ionosphere, especially the substorm cycle of the magnetosphere. Most existing models of these phenomena are based on MHD, which is a natural starting point as it offers the simplest representation of large-scale plasma phenomena. But it is clear that space-weather forecasting needs to take account of plasma effects that are not handled by MHD, which include:

a. small-scale features that have crucial influences on the chain of physics linking the Sun to the Earth, most obviously the self-consistent inclusion of magnetic reconnection as a key process at many points in the chain;
b. the lack of thermal equilibrium across much of that chain of physics. The plasmas in the solar wind or the magnetosphere are never in thermal equilibrium as they are collisionless plasmas, so MHD cannot fully encompass their behaviour. In severe space weather, this can be a major issue where a significant fraction of the plasma pressure can arise from suprathermal particles.

In the near term, one key priority for space-weather research is improved modelling of the solar wind and magnetosphere, since this has widespread application, including the threat to power grids. Other near-term opportunities focus particularly on the generation and propagation of high-energy particles, and incorporating these processes into models of solar-radiation storms and the killer electrons in Earth's radiation belt, as well as models of auroral electron precipitation—which is a significant cause of surface charging on spacecraft in polar low Earth orbits.

Longer term, researchers are keen to address areas where we currently have only a limited understanding of the science, but where major advances in physics-based modelling could have significant potential. A key example is developing a better understanding of how we can detect magnetic activity developing inside the Sun, how this then leads to CMEs and solar flares, and most importantly how to integrate this understanding into the chain of physics that drives space-weather impacts at Earth. There is already much work in progress in the first two areas; the challenge for the space-weather community is to make the connections across the full chain of events. This integration almost certainly needs to be done statistically, reflecting the fact that solar eruptive activity exhibits self-organized criticality. Space weather therefore needs models that can provide probabilistic forecasts on the time and scale of eruptions, in particular the likelihood that a fast CME will be launched in a direction that could bring it close to Earth. Solar-wind propagation models can then be used to explore whether that CME will actually reach the Earth.

Thus there is considerable scope for future research that advances our ability to deliver accurate and useful space weather services. But the momentum of that research will be sustained only if we maintain and improve wider knowledge of space-weather risks, not least by ensuring that policymakers and the general public are aware of research progress and how it can be exploited to protect societies and economies around the world. This is not a job that can be done once, as there are always more people with whom the scientific community can engage. In particular, there is a steady turnover of people in decision-making positions, and we need to raise their awareness so that we can promote evidence-based policy on space weather, both in government and industry. This engagement is a vital element in supporting future research, not least to emphasize the scientific evidence on the real risk from severe space weather. Since there has been a dearth of severe, or even moderate, space-weather events over the past decade, there is now a lack of practical experience in many sectors at risk from space weather—which now needs to be balanced with broader scientific evidence.

Finally, no look into the future would be complete without some discussion on how space-weather risks will develop as technology continues to evolve. One piece of good news is that disruption to satellite communications may be reduced through a strong trend towards using frequencies higher than 4 GHz to increase bandwidth, which as a side effect will also eliminate most of the effects of ionospheric scintillation. At the same time, a number of emerging applications may be at risk from space weather, such as the likely growth in systems that embed the use of satellite navigation, such as driverless cars, road tolling, and control of rail systems. As already discussed, space weather could cause these systems to experience position errors due to unexpected changes in ionospheric delay, and loss of signal due to ionospheric scintillation. It is therefore crucial to raise awareness of space weather in the communities developing these applications, and of how adverse impacts can be mitigated by a mixture of good design, forecasting and nowcasting.