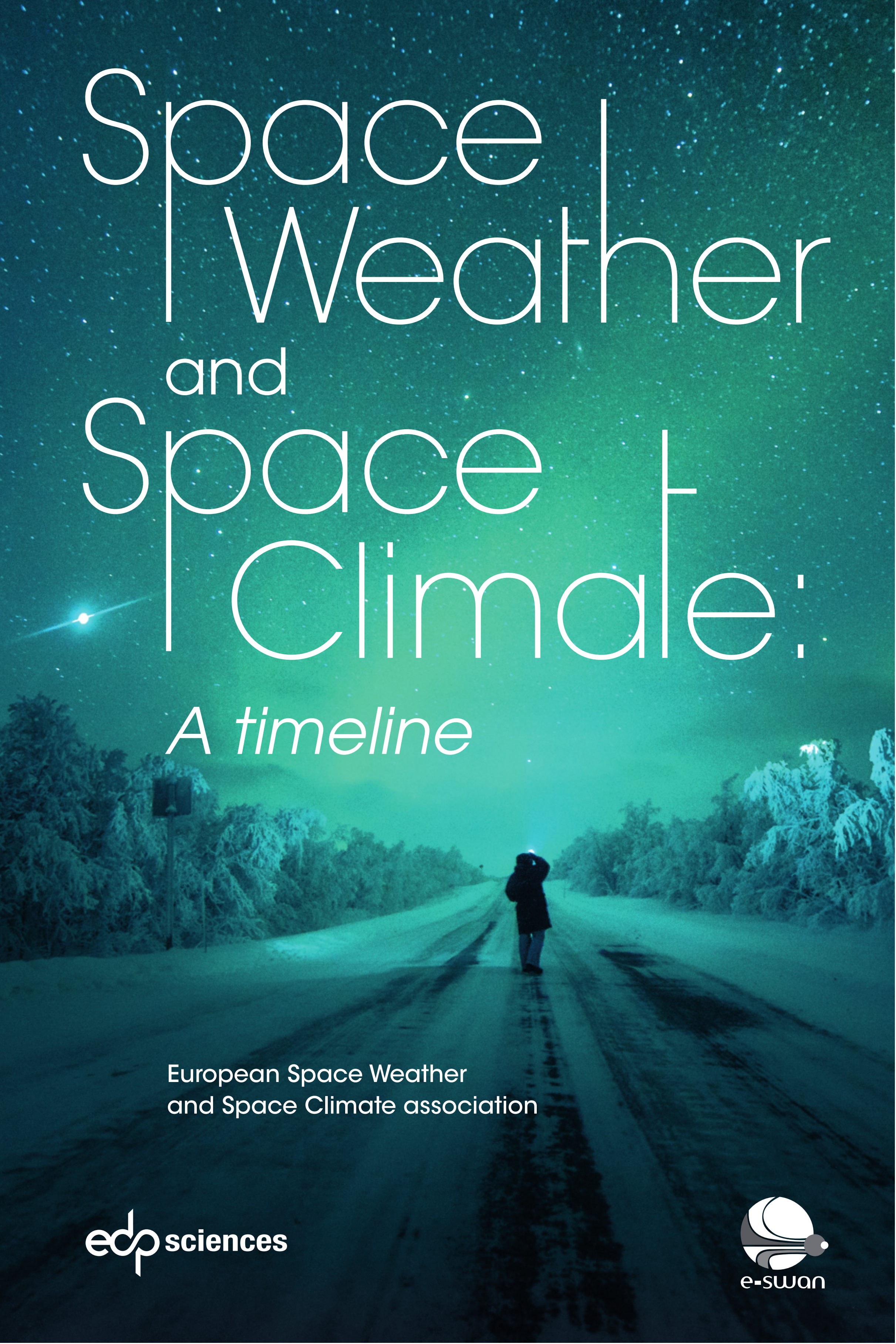


Space Weather and Space Climate:

A timeline

European Space Weather
and Space Climate association

edp sciences

e-swan

Space Weather and Space Climate:

A timeline

European Space Weather
and Space Climate association

At the beginning of the 20th century, it was admitted that our atmosphere could not extend beyond a few tens of kilometres. Shortly after the midst of that same century, it was understood that the Earth is surrounded by a complex and rich plasma environment, conditioned by solar activity, extending over several tens of thousands of kilometres. Today, we know that our planet's boundaries extend all the way to the Sun. This book tells the long story of humanity's efforts to understand the boundaries of the Earth and its influence from the external celestial objects dominated by our closest star, the Sun.

It also reveals the impact that solar activity is having on our technological societies, from the most beautiful to the most fearsome: Polar auroras and northern tourism, power and communications cuts, GNSS positioning degradations, planes losing contact with control towers, astronauts under threat, space… It presents, in a clear and educational way, this new and exciting discipline: space weather and space climate, its physics, its instruments, its methods, from modelling to artificial intelligence, its forecasting centres.

Written by a dozen of the world's leading specialists in this field, this book is the tribute of the world's largest space weather and space climate association, E-SWAN, to the lovers of space and nature.

edp sciences
www.edpsciences.org

ISBN : 978-2-7598-3610-9

9 782759 836109

e-swan
EUROPEAN SPACE WEATHER
AND SPACE CLIMATE ASSOCIATION

European Space Weather and Space
Climate Association

Space Weather
and Space Climate

A Timeline

edp sciences

Cover picture: credit Olivier Katz, AurorAlpes

Printed in France

EDP Sciences – ISBN(print): 978-2-7598-3610-9 – ISBN(ebook): 978-2-7598-3611-6
DOI: 10.1051/978-2-7598-3610-9

This book has been coordinated by Jean Lilensten and Jaroslav Urbar

The authors (in alphabetical order are): Sophie Chabanski, Antonio Guerrero Ortega, Marina Gruet, Magnar Gullikstad Johnsen, Stefan Hofmeister, Lauri Holappa, Jean Lilensten, Joao Pedro Marques, Lisa Nelson, Frédéric Pitout, Jaroslav Urbar, Christine Verbeke

Part of the current book is based on "J. Lilensten, F. Pitout, M. Gruet, J. Marques, Météorologie de l'espace, vivre demain avec notre soleil, Editions De Boeck supérieur, ISBN 978-2-8073-3306-2, 2021", with the authorization of the publisher De Bœck Supérieur.

Acknowledgements

This book is framed within the initiatives of the Outreach and Education Committee of the European Space Weather and Space Climate Association (E-SWAN). E-SWAN is an international non-profit association created to provide a long-term support to the space weather and space climate activities, with a focus on Europe.

The authors acknowledge the financial contribution of E-SWAN towards this publication via an International Coordination Action grant from the Research Foundation – Flanders (www.fwo.be).

J. Lilensten acknowledges funding support from the Programme National Soleil Terre (PNST – Solar Terrestrial physics national program) of the CNRS/INSU also co-funded by CNES and CEA.

Introduction

Space weather is a recent discipline, taking advantage of space physics at the interface between astrophysics and geophysics, but different in being between fundamental research and operational forecasting for industry. Space climate provides an understanding of the past, allowing us to predict variations in the space environment mainly due to solar activity; to quantify their effects on Earth and our technological world. In 2011, the OECD ranked space weather among the five significant global risks, equal to systemic financial risks, cyber risks, social unrest, and pandemics.

However, among them, space weather is still certainly the least known by the general public. The purpose of this book is to overcome this ignorance, not to add stress to an already highly anxiety-provoking world, but to inform and share. During the last decades, we have considerably pushed back the limits of our environment reaching out to the Moon and planets of the Solar System. We have pushed them back to the Sun itself. And we have learned that the Sun, in many ways, affects our life, environment, and technology on Earth. In what ways? Answering this question is one of the main goals of space weather and space climate, and constitutes the heart of chapter five of this book.

But how did we come to understand that our planet was bathed in the atmosphere of the Sun? By a long process.

Until the end of the 19th century, it was believed that the energy source of the Sun lay in chemistry similar to the burning of coal. Its gravity is also so great that scientists thought it impossible for anything to escape from it. As far as the Earth is concerned, the physics knowledge of that time explained well that the centrifugal effect compensated for the force of gravity at a few kilometres of altitude, above which there could no longer be an atmosphere accommodating "burned" material, therefore an ether was introduced on which "leaned" the light to propagate, with characteristics little constrained or known. However, questions appeared here and there. The first high-altitude soundings contradicted the predictions of such a thin atmosphere for the case of Earth, and the determination of the altitude of the polar auroras showed that there was something dynamic above 100 km. What could it be? The advances in electromagnetism also posed important problems. What was observed in the laboratory had to be reproduced in the Earth's magnetic field.

DOI: 10.1051/978-2-7598-3610-9.c901

How, and with what periodicity? The observation of sunspots showed a curious periodicity that seemed to correlate with that of the polar auroras: coincidence or causality? It was then the time of questioning.

But at the beginning of the 20th century, physics made considerable progress, and the nuclear energy source of the Sun was soon revealed. At the same time, radio transmitters revealed that our atmosphere extends far beyond the predicted few kilometres up to more than three hundred kilometres, where instead of electrically neutral gas it is mostly constituted by ions and electrons. Where do the ions and electrons come from? Adventurous physicists suggested that the main source is the Sun. It took an incredible amount of audacity for them to assume that particles could escape from it despite its fantastic gravity. The real-time of adventure started with the discoveries of the space environment, the topic of chapter two.

The space age and the increasing complexity of the instrumentation on the ground soon contributed to the picture in all its richness providing a multitude of details. The rigid and static picture that prevailed until the 1950's was put in motion showing how much the space environment can vary from one hour to the next one. Everything moves and everything mixes the particles with the electric and the magnetic field. The zones that surround us are manifold, subtle, always interacting, and never clearly delimited. Chapters three and four give pride to this conceptual revolution, the latter focused on the understanding of the Sun.

The sixth and last chapter is about the future of the discipline. Still, experience has taught us to be cautious, and we know that even if we write about the future, the future is not written. It is the opening towards modern methods still in a state of clearing for operational forecasts: networking, big data, and artificial intelligence. Thus, space weather and space climate have transformed from bench research to the most sophisticated operational applications within one century.

The reader will certainly notice a detail that surprised us during the writing of this book. Chapters one and two are filled with scientists' names. It is exciting to see how, long before the digital age, there were already lively and abundant exchanges in science, from one country to another, without language barriers. Then, subtly, chapters three and four – the time of complexity – substitute space probes and large terrestrial instruments for humans. The names of the scientists disappear, we can only guess that they continue to exist behind the instruments, but we see that they have almost become secondary, erased by a torrent of data. Chapter six is the logical continuation of this evolution: big data and artificial intelligence take precedence this time over the big instruments themselves. A single scientist can no longer master the calculations; entire teams are needed, where the individual seems to be diluted in the mass.

While it is difficult to escape this trend, we think that it is an illusion. The authors are all involved in space weather and space climate research, working on the ionosphere, the magnetosphere, the aurora, the Sun and its influence in the heliosphere, artificial intelligence, and operational services. We meet our colleagues during meetings and measurement campaigns. We know that space weather and space climate are above all driven by humans, carried by nature enthusiasts, and scientists eager to understand and predict.

By the end of this book, the readers will have gone through four centuries of science history. We hope that they will have understood how dependent our technological society is on a space environment firmly anchored to solar activity, and that a new hazard is to be considered for the survival of humanity. Not to scare, but to understand and prepare.

Contents

Chapter 1

The Age of Questioning

Space has naturally been a source of scientific inquiry since time immemorial. Long before the seventeenth century, legitimate questions, based on reliable observations, were raised. The seventeenth century saw the emergence of mathematics in the natural sciences, and with them, the possibility of predicting phenomena. Measuring instruments allowed new explorations. But each unveiling generated new questions which accumulated... They accumulated for several centuries. Where were we at the beginning of the twentieth century (a century that will see so many upheavals)?

Electricity and Magnetism

Perhaps what makes space weather different from traditional terrestrial weather is its strong and explicit dependence on the laws of electromagnetism. To introduce the beautiful history of space weather, we must start with the development of electricity and magnetism. It is curious that the science of electricity did not start to develop until scientists were able to reproduce conditions of outer space, the vacuum, in the lab, in the 17th century.

But electricity is not a recent discovery. Its manifestations had been noted since antiquity; Egyptians and Arabs related discharges from fishes to electric effects and Egyptians are known to have used high poles covered with copper to protect them from “what came from the sky”.

In the 7th – 6th century BCE, the Greek scholar Thales of Miletus (c. 624/620 BCE – c. 548/545 BCE) noted that amber acquires, by friction, the property of attracting light bodies. We can see this effect in our daily life with the use of artificial textile clothes such as fleece, made of polyester. The rubbing they undergo when we wear them is enough to charge them with electricity (provided they are dry). After that, they easily attract light objects, and can even produce real sparks when we take them off.

Thales also noticed the power of the magnetic force in the attraction of iron by lodestones without knowing that the Chinese and the Egyptians made the same observations long before him. On the properties of amber, he offered no explanation but on that of the magnetised stones, he suggested that effluents come out of both

DOI: 10.1051/978-2-7598-3610-9.c001

the stone and the iron. It will take 25 centuries to understand that the physics of the electric and magnetic phenomena are related together.

Later, it was understood that friction is only one of the ways to electrify materials, for example, something like heating could have the same effects. The ancient Indians noticed that certain crystals had the property of attracting hot ashes.

The Chinese, who were very advanced compared to other civilizations, invented a way to navigate thanks to magnetism two centuries before the Christian era, which soon led to the invention of the compass. This was a considerable step for humanity.

After these heroic ages where legend mingles with history, it will be necessary to cross a vast period to find the first scientific works on electricity and magnetism. The introduction of the compass in Europe in the twelfth century, perhaps thanks to the Arabs, gave rise to a very rigorous experimental study in 1269 by the military engineer, in the service of the Duke of Anjou, Peter Peregrinus of Maricourt (n.d.). In his study of the magnets, *Epistola de magnete*, he named for the first time the two ends of a magnet "north pole" and "south pole", producing opposite effects. He also made the confounding observation that if you break a magnet, each piece has in turn a north and a south pole.

The worldwide success of the word *electricity* is due to the English physicist of Queen Elizabeth I, William Gilbert (1544–1603) who took the word from the Greek word *elektron* which means Amber. In 1600, Gilbert published *De Magnete* describing the Earth as a huge magnet. He also showed that a permanent magnet could magnetise iron. Then he found in various insulating substances, such as glass or resin, properties identical to those of amber. He made the first measurements concerning electrical charges and recognized that there are bodies that conduct electricity well, which are good *conductors*, and others that oppose the passage of a current, *insulators*. In the first category, we find metals, such as copper, silver, and aluminium. In the second category, we can mention wood or air, the latter one being such a poor conductor that potential differences of hundreds of thousands of volts are needed to create lightning discharges. All this led him to compare the two forces, electric and magnetic, without having the physical means to recognize their similarity. However, his work will have a particular impact in the context of the Sun-Earth relationship. Indeed, it strongly impressed the astronomer Johannes Kepler (1571–1630), who drew from it a confirmation of the existence of attraction at a distance. Kepler suggested, at the beginning of the seventeenth century, that the force between the Sun and the planets could be magnetic in nature. Newtonian physics swept away these hypotheses, and it was only in the second half of the twentieth century that it was recognized that they were not totally unfounded, although through phenomena of which Kepler had no knowledge.

In 1672, German physicist Otto von Guericke (1602–1686) made a sphere of sulphur turn rapidly on a woollen cloth. Then, improving his device, he invented the first electrostatic generator. Guericke was also interested in the properties of the vacuum, which is of course not indifferent to the space environment. In an experiment that has remained famous throughout the world, he had two large hollow half-spheres built and placed against each other. Then, after having made the vacuum inside, he attaches to them 16 horses on both sides. Pulling with all their strength, the horses fail to separate the half-spheres. Indeed, these undergo the atmospheric pressure that no interior pressure compensates for. This experiment, which took place on the public square of Magdeburg, amazed the onlookers and the city authorities.

But let us return to electricity. A little later, the Englishman Stephen Gray (1670–1736) studied the phenomena of conductivity and electrification. He became one of the founders of that part of physics known as *electrostatics*, which studies electrical charges in their static configurations but not the action of currents.

The apparent existence of two types of electricity, as demonstrated in 1733 by the Frenchman Charles-François de Disternay du Fay (1698–1739), known as Dufay, was a real enigma. For example, one is attracted by an elder tree stick while the other is repulsed. It was not known at the time that these two types of electricity, called *resinous* and *vitreous*, represent two aspects of a single reality.

Then, in 1745, the Dutchman Van Musschenbroek (1692–1761) invented the *Leyden jar* with two other scientists. It is a glass bottle with a metallic element inside and coated externally with tin or aluminium. The glass is the insulator, and the metal plates are the two electrodes of our modern capacitors. By connecting these electrodes to a source of charge, the bottle capacitor is able to store a certain quantity of electricity. If then a reckless experimenter touches these two electrodes simultaneously with both hands, the bottle is discharged through the body giving back the stored charge with a nice jolt.

Benjamin Franklin (1706–1790) renamed Du Fay's electrics "*positive*" to designate the vitreous, and *negative* for the resinous. He also demonstrated that the sharp bodies are particularly effective to receive the electricity as to emit it. He thus invented the lightning rod and proved the electrical nature of lightning.

These experiments marked the end of a long preliminary experimental phase; they were followed by quantitative studies that made the most of precise measurements that would allow the general laws of electricity and magnetism to be identified. Two hypotheses were then confronted. That of the two fluids followed the experiments of Dufay. The one of clean, discontinuous electric matter came from Franklin's observations.

The physical interpretation of electricity and magnetism will resume its progression with the Englishman Henry Cavendish (1737–1810) and the Frenchman Charles-Augustin de Coulomb (1736–1806). Cavendish studied the properties of the electric force, and Coulomb established its fundamental law. Let's stop for a moment to admire this real tour de force. In the wake of Isaac Newton (1642–1727), we pass a stroke of genius from observation to its mathematical formulation, opening the door to modern physics. This conceptual leap marked the birth of our contemporary world.

After quantifying the electric charge, Coulomb stated that the force between two electric charges was proportional to their product and inversely proportional to the square of the distance between them. These discoveries plunged physicists into an abyss of perplexity because Newton's law of universal gravitation was very similar to Coulomb's. There, the masses play the role of charges. In both cases, a constant of proportionality was needed. Was the universe governed by a very small number of physical laws?

Based on these lessons, Hans Christian Ørsted (1777–1861) in Denmark realised that electricity could produce heat, and light and act on the chemical composition of bodies. He naturally wondered whether it could be identified with magnetism. To do this, he carried out a simple experiment that all physics students in the world have reproduced during their studies: he placed a magnetised needle parallel to a conducting wire in which an electric current was flowing. The needle was more or less attracted by the wire depending on the intensity of the current, and the pole attracted by the wire depends on the direction in which the current flows. So, there could be an electrical influence on a needle that was supposed to react to magnetism. Moreover, this influence occurred at a distance, without contact between the electricity and the compass needle. Shortly thereafter, Henrik Lorentz (1853–1928), in the Netherlands completed this approach and demonstrated that if a conducting wire is placed in a variable magnetic field, an electric current is produced! This phenomenon is called *magnetic induction* since the magnetic field induces a current.

The concept of *field* was another intellectual revolution, first established in 1855 by the British scientist James Clerk Maxwell (1831–1879) it is based on the idea of expanding the properties of some source, the charge in this case, into space, in what we call the region of influence. In the context of the space environment, we will mention the magnetic field, the electric field, and the gravitational field. Maxwell's fundamental contribution was to unify the work of previous prestigious physicists through the concept of field. To do this, he managed to unify the physics of electricity and magnetism in over twenty equations that later were synthesised by Oliver Heaviside (1850–1925) into only four equations that now bear Maxwell's name.

In two of these equations only one field appears, either the electric field or the magnetic field. The first one (Maxwell–Gauss, or Poisson equation) says that the electric field is directly derived from the density of electric charges in the medium, that is to say, the density of charges of negative electricity–electrons[1] – or of positive electricity–ions -. The second Maxwell's equation (sometimes known as Gauss law for the magnetic field) says that if there are particles that carry electric charge, there are no particles that carry magnetic charge. These equations do not solve the problem of coupling between the two fields. For this, two additional equations must be considered in which both fields appear on each equation. The Maxwell-Ampere equation is explicit. What it states is that a magnetic field appears when a movement is set up. This movement can be represented by an electric current that is to say, a displacement of charge, or a time variation of the electric field for one reason or another. The Maxwell–Faraday equation indicates that in return, a variation of the magnetic field over time or relative to space can generate an electric field.

These four equations tell the physicist much more. In particular, they tell us in which directions the fields are created. For example, if the charged particles spiral, they create a magnetic field at the centre of the spiral along the axis of rotation. They also tell us how they propagate: by waves, and at the speed of light. What

---

[1] The electron was not discovered until 1896 by the physicist Joseph John Thomson (1856–1940), who received a nobel prize for his contribution the conduction of electricity through gases, a topic intimately coupled with space physics.

Fresnel sensed, in advance of Maxwell's confirmation, is that light is only a particular case of electromagnetic waves. But the important, fundamental point is this: an electric charge creates an electric field. It is when the charge is in motion that the *observer* sees a magnetic field superimposed on the electric field. An observer moving with an electron would only see the electric field. One must see the electron passing by to feel the effect of the magnetic field.

The other extraordinary observation is that if we apply an electric field on charged particles, they are disturbed and start moving in the direction of the field; if we apply a magnetic field on moving charged particles, they are deviated perpendicularly to both, the field and the direction of movement. This is how the electromagnetic forces are created, written in another equation called the Lorentz force equation which together with Maxwell's equations complete the Electromagnetic picture.

We can then reanalyze all the past experiments according to the Maxwellian theory. Why does the Earth have a magnetic field to which compasses react? Because in its outer core, in what is now called the liquid core, electrically charged iron and nickel move slowly in a spiral, creating a planetary magnetic field. Why do two magnets appear when one is broken? Because in some elements, especially metallic ones, electrons are spinning in the depths of matter. As they spin, the electrons generate a small local magnetic field, a *magnetic moment.* Generally, in non-magnetized substances, the rotations of all the electrons take place in any plane with random directions, and so the magnetic moments cancel each other out. But in a magnet, all electrons rotate in the same plane and in the same direction. The magnetic moments add up to a macroscopic magnet. Breaking the magnet does not change the direction of rotation of the electrons. To stop the magnetization, energy must be supplied to the electrons by rubbing or heating so that they undergo a vibration stronger than the rotation.

Thanks to Maxwell, we discovered that it is difficult to separate fields and particles. In the rest of this book, we will talk about the *solar wind* composed of electrons and protons, but also sometimes about the magnetic field that is associated with it, the *interplanetary magnetic field.* Depending on whether we wish to examine one property or the other, we will look at the equations of the particles or those of the field, but at the end, we will be looking into Maxwell's equations.

The other concept is that of line of force, put forward by Faraday. Let's go back to Oersted's experiment. Since we now know that electricity and magnetism are one and the same, *electromagnetism*, let's replace the wire with a magnet. This way, you can do the experiment at home. Simply move the compass around the magnet. When it is in front of the + pole, it points to its − pole, and *vice versa.* But as it passes from one extreme point to the other, the needle describes a nice, rounded curve, which Faraday explains and calls a *field line.* Thus, when a magnetic field is applied to moving charged particles, they are not deviated in any direction: they will rotate around the lines of force of the magnetic field. They can move forward or backwards, depending on whether they had an initial forward or backward motion, but they do so in a spiral. The direction of the spiral depends on the sign of the electric charge, it is the opposite for positive and negative charges.

We have almost finished with the physical basis necessary to understand our space environment. Theoretically established by Maxwell, the existence of electromagnetic waves had yet to be proven experimentally. This will be the work of the German physicist Heinrich Hertz (1857–1894) in 1887. Thus, these waves in which we literally bathe bear his sweet name: Hertzian waves.

For this, Hertz imagined the following device: between two small very close spheres, he made sparks fly; the equipment used for this purpose is called a *spark gap*. These sparks are accompanied by waves that are detached and travel away from their sources being able to act in an open metal loop located nearby; if they are sufficiently intense, they generate a detectable voltage (said to be *induced*) in the loop, which in turn could cause sparks to fly between the open ends of the loop. However, such a reception device is not very sensitive and, to better highlight the Hertzian waves, the Frenchman Edouard Branly (1844–1940) invented the “metal filings coherer”: he noticed that the iron filings agglomerated when the waves were manifested and that then, its resistance to the passage of the current varied. Thus, he conceived the penultimate link of a chain whose last link, the invention of the antenna, would lead to the birth of the *radio* and the *radar*. With these instruments, we will no longer need telephone wires to communicate at a distance because from now on, there are waves that travel in space at the speed of light and ensure the link between the transmitter and the receiver.

The first radio link was not the work of Branly, but that of a Russian, Alexander Popov (1859–1906). Popov carried out the first experiment which marked the birth of radio, on May 7, 1895: at the school of Torpedo boats in Kronstadt, he developed a complete device for receiving Hertzian waves.

Then, on March 24, 1896, in the presence of the members of the Russian Society of Physics, Popov proceeded to an incredible experiment of remote communication: the transmitter was in the buildings of the Institute of Chemistry of the University of Saint Petersburg. The receiver was some 250 m away, in the room of an old physics cabinet. The transmitter was put into service and, in front of a dumbfounded audience, the receiver came to life and communicated to a paper recorder the Morse signals it received. These signals spelt out the letters of a name: Heinrich Hertz, in homage to this precursor. The success was total. The successes accumulated: the first radio link between the Eiffel Tower and the Pantheon by Eugene Ducretet (1844–1915) in 1898, the trans-Channel message on March 28, 1899, by the Italian Guglielmo Marconi (1874–1937), the creation of Wireless Telegraphy, or TSF, the forerunner of our radios in 1903 by Augusto Righi-Dessau (1850–1920), who named it and Gustave Ferrié (1868–1932), who created the Eiffel Tower station.

Almost at the same time, in Great Britain, Ernest Rutherford (1871–1937), born in New Zealand, did a similar experiment on more than three kilometres.

As for the radar… It will be a story so intertwined with that of the discoveries of our space environment that we will dedicate a future paragraph. Let’s rather see now the atmospheric works which took place in parallel to the discoveries of electromagnetism…

The Atmosphere and the Aurora

At the dawn of the 19th century, the accepted picture of the atmosphere was that of a thin layer of gas, a few kilometres thick. It was already modelled thanks to the works of thermodynamics. We knew that the pressure decreases with temperature. It had also been measured that the temperature decreases as we go up, and several laws had been proposed. Pierre-Simon Laplace (1749–1827) used recent advances due to Robert Boyle (1627–1691), Edme Mariotte (1620–1684), and other scientists, and proposed the ideal gas law, which links pressure, temperature, volume, and density of a gas through a mysterious number called the ideal gas constant, and which will be fully explained at the end of the 19th century. Another law, that of dynamics, completes it. It expresses that the variation of the pressure with the altitude is inversely proportional to gravity, to the density of the gas, and the thickness of the gas. With the help of these two laws, Laplace could theoretically determine how the atmospheric pressure and density vary with altitude. He discovered that this variation is exponential: the higher we go, the faster the pressure decreases. However, it was not possible for him to demonstrate how temperature varies as it does (we know today that we must consider the composition of the atmosphere and in particular its water vapour content). By means of measurements, Laplace proposed a relatively simple empirical law to describe the variation of temperature. According to this law, the temperature decreases as a function of the square root of the altitude. But this law was very approximate and soon it was simplified into a linear version.

Innumerable experimental verifications were carried out at that time which all confirmed Laplace's law up to altitudes of 25 km. For this purpose, balloons were used, and, from the beginning of the 20th century, the development of radio transmissions allowed them to carry instruments (meteorographs) and send the data from the balloon down to the ground in real-time (radiosonde). The difficulty of airborne measurements and the determination of their variation with altitude was well known: the determination of the altitude alone was a feat. It was naturally out of the question to evaluate it according to the pressure since it was not known how it changes with altitude. Theodolites, instruments to measure angles, were used to measure the elevation. But to measure the altitude, it was necessary that the balloon was in sight of at least 2 theodolites, and if possible 3. But as the balloon, carried by the wind, did not remain stationary, it was necessary to set up networks of theodolites, more than ten, with fast operators to handle each of them. The notes of these operators had to be made exactly at the same time, otherwise the determination of the altitude would have been impossible. The measurements of elevation of the balloon were thus reported every quarter of an hour and the operator in the balloon gondola had to take also at those exact times a photograph vertically pointing down to the ground. The comparison of the position in the photograph and that calculated from theodolite recordings made it possible to deduce the altitude.

At the beginning of the 20th century, we witnessed a real revolution in the conception of the atmosphere: for several years, Teisserenc de Bort (1855–1913) explored the atmosphere with probes embarked on kites, the *meteorographs*. His instruments measured temperature, humidity, and pressure. They managed to verify the law of Laplace. Cleverly taking advantage of all the technical progress, he embarked his instruments on balloon probes perfected in 1898. These allowed him to overcome the barrier of ten kilometres and to explore staggering altitudes of 30 km, reaching later

35 km. In 1902, his observations overturned all predictions about our environment: he discovered that while the pressure continued to decrease, the temperature remained constant. We say that the layer is *isothermal.* Teisserenc named it the *stratosphere.* How to explain this new puzzle? And how far does the temperature remain constant? What physical process maintains enough energy exchange to heat this already rarefied gas to minus sixty degrees Celsius? Have we found the ether, the hypothetical fluid that physicists have long predicted? An intense activity was developed to explore this new frontier. It was discovered that at the lower boundary, called the *tropopause*, it varies according to the latitude and the season. They realised that this is also the case for the temperature of the stratosphere itself. Multiplying the experiments in many geographical conditions and seasons, they realised that these amplitudes of variations are higher at high latitudes.

The succession of measurements showed that the temperature was not, however, rigorously constant with increasing altitude. Sometimes it seemed to decrease again but sometimes, on the contrary, it rose slightly with altitude.

Whatever the mechanism, the representation of the Earth's atmosphere at the dawn of the 20th century was still relatively simple. The ether was only pushed a little higher, over two layers instead of one. However, one fact kept observers worried since antiquity, which is the sporadic appearance of polar auroras. The observation of these atmospheric lights had already a long history: Aristotle compared them, in his book I of Meteorology to "a flame mixed with smoke … and to the blaze of a meadow whose stubble is burned". People living at high latitudes feared them. There, they were thought to bring death. It was not until the sixth century that we find the first description of it without mysticism, and it is due to the Bishop Gregory of Tour (538–594) that we owe it. He testified to having seen "brilliant rays of light that seemed to collide and to cross each other, after which they separated and faded away… If it had not been night, one would have thought one was seeing the dawn". Taking up his phraseology, a French astronomer, philosopher and theologian, Pierre Gassendi (1592–1655) named, in 1621, these lights *aurora borealis* (boreal, because seen always towards the north and aurora from the goddess of dawn), a term also used in several other languages.[2] However, there is debate as to whether the authorship for introducing this name should be transferred to Galileo in 1619. Aurora's appearances were few and far at mid-latitudes and always aroused great emotions. The registered examples are numerous; thus, in 1583, between 800 and

---

[2] Here are some translations: in Armenian, Husisapail means "northern lights". In Finnish, revontulet (accentuate the "r" at the beginning of the word) means "the fire of the fox". In Dutch, "Poollicht", naturally means "polar light". For the north, the Dutch also speak of "noorderlicht". The Flemish also have "Zuiderlicht" for the south but more commonly use "Poollicht" or "Aurora". In Italian, we also speak of polar aurora: "aurora polare" (singular) and "aurora polari" (plural) which, like in French, are boreale/boreali in the north and australe/australi in the south. In Greek, Πολικό Σέλας is pronounced "polikó sélas" and means "polar aurora". In Ukrainian, полярне сяйво is pronounced "pol'arne" with the "l" as a soft consonant and then "s aivo" or "s-ai-ivo" (soft s and the very short, almost consonant i). The literal translation is "polar light". Germans and Austrians say "Polarlicht" and "Nordlicht". The Anglo-Saxons use the term 'Polar lights,' but they are also occasionally referred to as 'auroras,' which may indicate Latin influence. In Slovenian, they say "polarni sij", 'J' being pronounced as aii. This can be translated as "polar flash".

900 people in France travelled from La Ferté-Gaucher to Paris to gather in the Great Church, because they had seen "signs in the sky and fires in the air". This was still the case in 1938 when a red sky (typical of mid-latitude auroras) was seen in many European countries as a sign of a great fire or, in the case of Spain, going through one of the most intense moments of the civil war, soldiers feared that the red sky was the result of the use of some new and powerful weapon.

Why did auroras baffle atmospheric physicists at that time? Because all the efforts made to determine their altitude showed that they were located very high. Did they occur inside the atmosphere or outside? Were they an optical illusion, a reflection of the Sun, a reflection of the Earth?

To understand – and accompany – the journey of these pioneers, let us look at the conquest of the polar areas where the aurora was more frequent.

In the year 1579, the Dutch, who had just gained their independence, had been chased from their trading post in Arkhangelsk by the English and were looking for new commercial outlets. The enthusiasm and the money of a French shipowner who immigrated to Holland, also a navigator, privateer, and merchant, Balthazar de Moucheron (1552–1630), won the decision of the Dutch government, and in June 1596, after two unsuccessful attempts, two ships embarked to explore a new maritime route to China in the northern seas. The chief pilot was William Barents (1550–1597), a remarkable navigator, who was familiar with the previous attempts and had himself already participated in two unsuccessful expeditions in 1594 and 1595. Heading straight north, the sailors passed the Norwegian coasts, crossed the latitudes until they discovered a small island that they named "Bear Island", then a large one on which they made an incursion while being convinced to be on a coast of Greenland. Covered with majestic peaks, they called it the *sharp mountains*, or in Dutch *Spitzberg*, ignoring that the Norse had already discovered it around the 10th century. But without regard for the cooling weather, Barents continued his exploration tirelessly instead of returning to Holland. He sailed to the east, providing his name to the sea he was the first to navigate. On August 17, 1596, "in terrifying cold and black misery", the 17 sailors began the first Arctic wintering in history in the bay of the "ice port" on the island of the "Novaya Zemlya". After a terrible polar night, they tried to escape in rowboats in June 1597, their ship being unusable. Barents died in the adventure, but the survivors, dressed in bear skins and wearing white fox furs, became the heroes of the first great Arctic epic, capturing the imagination of the world and especially that of the great maritime nations.

The Englishman Henry Hudson (1565–1611), record holder of high latitudes (80° 23′) in 1607, looked first for the North-East passage, then in 1609, decided to explore a hypothetical North-West passage. He gave his name to the famous bay of the American North and established a colony that would later become New York. His ship, named the Discovery, was the first large ship to explore the northern seas. On board, William Baffin (1584–1622) sailed a few years later up the west coast of Greenland and discovered Smith Sound. He gave his name to a sea and in 1821, Edward Parry (1790–1855) named Baffin a nearby island. On the idea of the scientist Johannes Werner (1468–1522), from Nuremberg, he measured for the first time precisely the longitudes by the process of "lunar distances". This process consists of using the measurement of the angular distance between the centre of the moon and a star or a planet and comparing it with tables of the same distance established in Greenwich, England.

In 1724, Tsar Peter the Great (1672–1725) summoned the Dane Vitus Bering (1681–1741) to St. Petersburg, asking him to draw a map of Siberia. He knew that whales carrying two Dutch harpoons had been found recently off the coast of Kamchatka, in the middle of the Pacific Ocean under the latitude of 60°. It was for everybody the proof that there was a west–east passage in the north of Russia. Bering crossed 8,000 km in Siberia by land before undertaking the construction of his ship, all materials needed having been taken with him on board sleds, an unheard-of exploit. He passed through the fog of "his" strait, which today separates Russia and the United States, before turning back. After these discoveries, a colossal expedition brought together a thousand men under Bering's leadership six years later. In this crowd were academics, astronomers, and geographers. Bering died exhausted on the island that now bears his name and which, on the Russian side, seems to close the American Aleutian Islands.

Soon enough other explorers attacked the Great South. In 1738, the Frenchman Bouvet de Lozier (1706–1788) saw the first tabular icebergs at 48°50′ South. He described their fauna as penguins and seals. His compatriot Yves Joseph de Kerguelen de Trémarec (1734–1797), who had already made a trip to Greenland, discovered "his" island at 49°40′ South. But it is to Captain James Cook (1728–1779), English navigator, that we owe in 1775 the recognition of the Antarctic, with a record at 71°10′ South.

The battle of Waterloo made Great Britain the first world power. Its navigators systematically explored all seas, with deliberate scientific ambitions. Records fall. William Scoresby Jr (1798–1857) took advantage of a mild summer to approach Greenland between 72° and 74°N. Then Parry left England in 1819, went around Greenland from the south, up its western coasts and turned north of Baffin Island. There, in Lancaster Sound, they noticed that the compass needle was going crazy because of the proximity of the magnetic pole. They had to steer without it. It was only at about 110° West longitude that Parry was stopped by ice. No matter, he had just found a passage to Canada. With this achievement and a remarkable polar wintering, Parry set out to conquer the North pole and found himself blocked at 82° 45′, 700 km from his goal. The struggle is fierce. Ross, in the south, discovered Possession Island and explored the shores of Antarctica that he dedicated to Queen Victoria in 1841. They tried to use steam, paddle wheels and balloons to explore the Arctic and to conquer its pole, which became an issue of considerable prestige. Many heroes lost their lives, in a long litany of failures.

All these explorers carefully noted their observations, including those of the aurora. In 1839, an expedition led by Auguste Bravais (1811–1863) wintered the corvette La Recherche in the vicinity of the North Cape to make the first modern auroral observations. In 1860, Elias Loomis (1811–1889), an American meteorologist from Yale University, published the first known map of the frequency of occurrence of the aurora borealis. In 1871, Hermann Fritz (1830–1883) drew a more precise map from the scattered notes of the explorers, showing the occurrence of night-time auroras up to low latitudes.

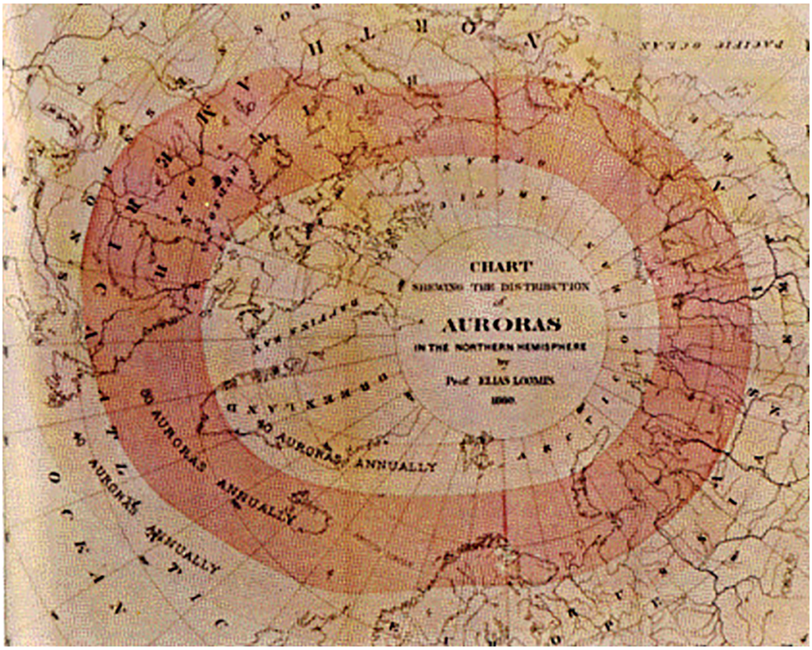


FIG. 1.1 – The aurora map by Loomis. The dark band corresponds to a frequency of at least 80 auroras per year.All rights reserved- © 1860 E. Loomis / www.phy6.org.

What emerged from this work was that the auroras appeared in a ring centred around the geomagnetic pole, the so-called auroral zone. It seemed that there was a direct link between the Earth's magnetism and these strange lights. The auroral zone should not be confused with the auroral oval. Indeed, the auroral oval, with its centre somewhat offset from the geomagnetic pole, was only confirmed by all-sky camera observations during the International Geophysical Year 1958–59. Friedrich Christoph Mayer (1697–1729) was the first to assume the centre of the auroral zone – ring around the geographic pole, as part of his work with altitude determination of the aurora at the beginning of the 18th century. His speculation was later further supported by the Swedish scientist, Wargetin (1717–1783), who concluded in 1752 that the ring would have its centre in a north-northwestern direction as seen from Europe. In 1827 Christoffer Hansteen (1784–1873) published, possibly, the first illustration of the auroral oval around the pole. Based on his observations during the Vega expedition through the North-East passage 1878–79, Adolf Erik Nordenskiöld (1831–1901) concluded that the aurora would be a ring with a centre somewhere between the geographic and geomagnetic pole.

The link between the appearance of auroras and magnetic disturbances was discovered by Swedish workers Hiorter (1696–1750) and Celsius (1701–1744) in the 1740s. After studying the direction of a magnetic needle for a whole year, they found the relationship that would forever unify the fields of geomagnetism and auroral science.

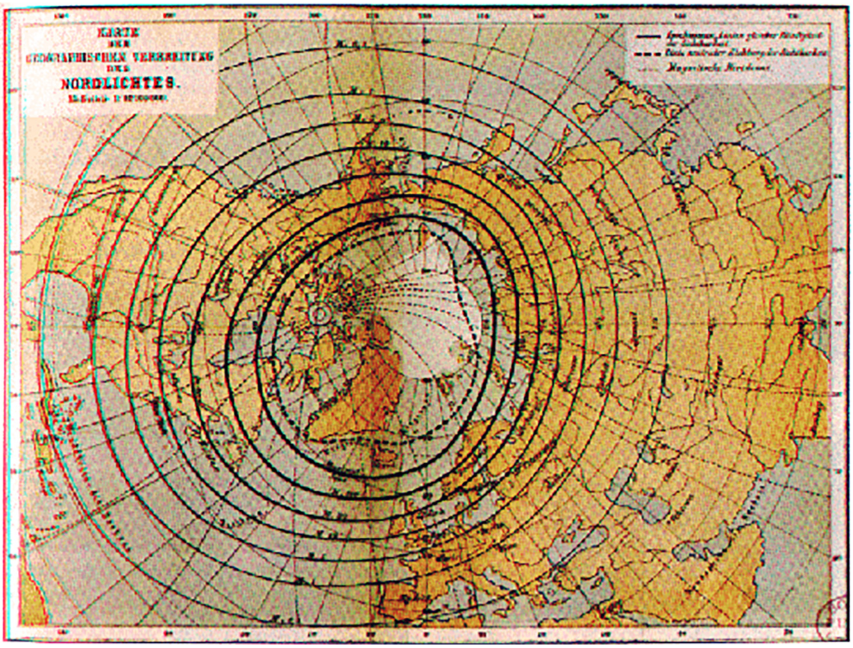


Fig. 1.2 – Map established by Fritz in 1881, according to data going from 1700 to 1872. One distinguishes there the lines of equal auroral frequency, or "isochasms" for the northern hemisphere. The most equatorial isochasm corresponds to 0.1 aurora per year, or 1 aurora every 10 years. Moving northward, we find successively the isochasms 1, 5, 10, 30, 100, and more than 100 or "maximum". Thus, we see that between Vienna and Brussels, the probability of observing auroras according to Fritz is 1 per year to 5 per year. We now know that it varies greatly depending on solar activity.

Alexander von Humboldt (1769–1859), a German nobleman, got interested in the Earth's magnetism and decided to do extensive journeys spanning from South and Central America over Europe to Central Asia to measure the geomagnetic field over the known world. Besides that, he discovered that the total intensity of the geomagnetic field increased with latitude; he became aware that the magnetic field was not stable over time, but that sometimes the intensity of the magnetic field seemed to vary dramatically over time. He named this phenomenon Magnetische Stürme, *i.e.*, magnetic storms and believed that the appearance of auroras and the vibrations of the magnetic needle had the same, common cause. Subsequently, he initiated that several magnetic observatories be placed over all Europe and Russia and that the intensity of the magnetic field shall be recorded at a synchronised, regular interval. From 1829 to 1835, about 20 magnetic observatories collaborated, and they showed that the magnetic storms did not appear the same at all locations, but that they are indeed a function of local time.

At this time, the idea of a scientific cooperation of all nations in polar exploration was born, at the instigation of the lieutenant of the Austrian Navy Karl Weyprecht

(1838–1881). It led in 1882–1883 to the first *international polar year (IPY).* It was the first time that eleven nations collaborated for essential scientific purposes – even if, of course, one should not underestimate the ulterior motives of each of them: until then, international collaborations were the work of the scientists themselves. Few great discoveries came out of it, but the decision was taken to renew the experiment every fifty years.

One man illuminated polar research. He was the Norwegian Fridtjof Nansen (1861–1930). He was a brilliant negotiator, it was thanks to him that his country obtained independence from Sweden, which was twenty times more powerful. The repatriates of the First World War were able to return home, thanks to him too, but also the White Russians who were wandering in Europe were able to obtain a new nationality. It was he who took care of the victims of the war between Turkey and Greece, and who managed to feed millions of people during the Russian famine of 1922. It was this great man who was awarded the Nobel Peace Prize that same year, the proceeds of which he distributed to charity.

But this wise politician was also gifted with a Viking physique and strength, a love of the wilderness and a solid culture. In his desire to conquer the pole, he designed a light and sturdy sledge, a waterproof tent, sleeping bags made of reindeer skin with a hood and a portable stove of remarkable ingenuity. With the help of this revolutionary equipment, and with five companions, he managed to cross Greenland on skis in 1888, passing through mountains at minus 40° at an altitude of 2,716 m. Then he designed his famous ship, the *Fram*, which in Norwegian means *forward.* A short and wide ship in the shape of a walnut shell, the Fram never let itself be trapped by the ice. Nansen wanted it to be able to "slide like an eel, out of the ice, when the blocks would grip it with force". Once his plan was conceived, it took nine years before it started to be executed, lasting three years for the construction of it. It is amusing to note that today, the design and construction of scientific space probes use exactly the same proportions! More than the other explorers, Nansen was passionate about science. He had noticed that the wrecks of the Jeannette, a ship that had sunk off the coast of New Siberia in 1881, had drifted towards the North West, past Franz Josef Land and Spitzbergen before reaching the waters of Greenland. His hypothesis was that the wood had floated north of Spitzbergen, on the pole side, and not south, on the European side. His crazy and brilliant project was to let the Fram drift and follow the path traced by the wreck before him. Leaving Siberia in June 1893, the ship brilliantly realised Nansen's prediction in almost exactly three years. Its designer had already arrived in Norway, because tired of this long drift, he had made an unsuccessful attempt to conquer the North pole, blocked at 86°13′6″ only 400 km from the goal, before going back down south risking death several times.

Following this first experience of international collaboration, 1901-1092 was declared the Antarctic Year. It marked the return of England to the scene of the conquest of the poles, thanks to the great explorer Robert Falcon Scott (1868–1912). With his ship, the "Discovery", Scott cleared more than 5,120 km of Antarctic coasts, bringing back a lot of scientific observations, and in particular those of the first photographic spectrometer for the observation of the southern lights.

It fell to the American Robert Peary (1856–1920), on April 6, 1909, to conquer the North pole, with four Inuit and his black servant. His strength was to have

Fig. 1.3 – The statue of Roald Amundsen after a snowstorm on the scientific base of Ny Alesund, at 78.8°N. On the left, the "blue house" where scientists from the French (IPEV) and German (AWI) polar institutes are staying (credit J. Lilensten, CNRS – IPAG).

studied for years the Inuit survival techniques, and in particular the pulling of the sledges by the dogs. Finally, it was Roald Amundsen (1872–1928), the Norwegian who in 1906 had found the North–West Sea passage during his quest to locate the magnetic north pole, who conquered the South pole in 1911, on December 14. Scott, his unfortunate rival, would follow him without the use of sledge dogs one month later, a negligence that he would pay for with his life in March 1912, during his return.

In 1929, Admiral Richard Byrd (1888–1957) dared to winter in Antarctica, by 78°35 S, filling his observations with the lack of data on aurora australis. He repeated his exploit, at the same time as a French mission in Terre Adélie in 1940.

Finally, in 1932–1933 the second International Polar Year took place, gathering this time the contribution of 26 nations. France's particular objective was the observation and understanding of the aurora. But in the meantime, many mysteries had been solved…

What do we know about auroras from the numerous observations of the early 20th century? They consist of more or less fleeting lights, of green, mauve, blue, yellowish, milky white or sometimes red colour. Their form is very variable, passing from curtains with dynamic volutes (draperies) to a uniform green coverage of the sky (diffuse auroras) while passing by fine stable and intense structures (the arcs). The luminous intensity is weak, at most of the order of a cloud lit by the full moon, that is to say, "a few micro-candles per square centimetre". Thus, the brightness of the full moon is enough to cover the auroral glow. It is also not possible to observe them when there is sunlight, so there was actually no evidence that they take place in summer in the northern hemisphere or in winter in the southern hemisphere during the so-called polar day. It was therefore perfectly possible that they had a link with the terrestrial seasons. In the north, they are concentrated on the zone highlighted first by Loomis, then especially by Fritz in 1881 as the enclosed maps depict. The work of the latter was completed in 1942 after finding that the auroras are magnetically linked to the Earth. Therefore, it was decided to refer to magnetic rather than geographical coordinates. How was this done? By creating a system of coordinates which, instead of passing through the geographical north and south poles, pass through the magnetic north and south poles. As these are not exactly aligned, we use "ideal" geomagnetic poles which would be those of a perfect magnet as defined by Gauss. Thus, we find that the auroras are particularly numerous at 70° north magnetic latitude. For the southern hemisphere, the lack of observations does not allow scientists to be conclusive, but it seems that the peak is at 72° latitude.

Auroras take place above the clouds, a fact attested by all observers. But at what altitude? Gassendi tried to determine it for the first time in 1621, followed in the 18th century by De Mairan and then by Bravais and Lottin during their 1838 expedition. Their measurements agreed with most of those made at the beginning of the twentieth century: between 80 and 200 km. However, Loomis observed them up to 1,000 km! These measurements are certainly flawed, but not to the point of confusing these high altitudes with the lower atmosphere! How do we proceed? First, by purely visual observation: two distant observers draw on a map of the sky the contours of an aurora at a precise time agreed between them. By triangulation, they then calculated the altitude. As soon as photographs were available, they were used to replace drawings. However, progress was not very fast because of the weak luminosity of the phenomena and their great dynamics. It was necessary to ensure that at least one identical star appeared on two distant photographs, a procedure employed by Carl Størmer (1874–1957) from the 1910s onwards. In 1946, this author was able to publish the results of 12,330 auroras (who would examine today at least 24,660 images by eye in order to make 12,330 parallax calculations?) He found 58 cases between 70 and 80 km of altitude, 1,280 between 90 and 100 km, 1,808 between 100 and 110 and only 3 between 1,050 and 1,100 km.

Others, such as Lars Vegard (1880–1963), focused on measuring the intensity of auroras as a function of altitude, with very different results. Their variation over time was studied and it was found that it depended on the place where the observation was made. When we trace the direction of the auroral arcs, we find more or less the map of Fritz.

From 1869, Anders Jonas Angström (1814–1874) showed an interest in studying in detail the light of the aurora by decomposing it through a prism. Each colour is the signature of a particular chemical element as shown by the work of Newton at the end of the seventeenth century, the astronomer William Herschel (1738–1822), to Joseph von Fraunhofer (1787–1826) in 1821.

Unfortunately, it was for a long time impossible to determine by laboratory measurements which element produces a given line of light. From 1912, Vegard recorded a considerable number of auroral spectra. He found that the most intense is the green line at 557.7 nm. The sensitivity of photographic plates varies with the colour observed so that at that time, it was still impossible to make numerical comparisons between lines. Vegard nevertheless showed the existence of several intense visible lines while other researchers examined the invisible radiation, in the ultraviolet and infrared.

Babcock (1843–1931) was the first to find an interpretation for the green line. It is the radiation produced by the atomic oxygen when it is excited in a certain way (called 1'S) following, for example, a collision, and then de-excites towards another state (called 1'D). However, this line is most often observed around 100 km in altitude. Therefore, because there is oxygen at this height, the aurora is an atmospheric phenomenon. In the fifties, another intense line of the aurora, the red line at 630 nm, was also identified. Once again, it is the atomic oxygen that is responsible, when it passes from the excited 1D state to its stable or fundamental state (called 3P). This red glow appears at around 250 km of altitude. At higher altitudes, lines of helium and hydrogen have been identified but also below, at around 100 km, molecular oxygen, and molecular nitrogen had been also identified.

Not only the so-called precipitation of energetic electrons cause aurora, also precipitating protons do. The so-called proton aurora, emitting Hydrogen lines, is different from the electrons, in that the protons themselves emit light, rather than the gas they collide with. A proton will spiral down the local magnetic field line, collide with the ambient electrons and "grab" one, thereby losing its charge. Then the magnetic field will lose its influence over the proton and this one will continue in the direction it had before the collision while emitting light in its path. Then another collision occurs, losing the electron, and again it will spiral downwards. This process will happen again and again, and protons will move sideways across the sky, creating a diffuse faint aurora covering most of the sky. The Hydrogen lines in the aurora were first observed by the above-mentioned Vegard in 1939 from Oslo and confirmed by the Auroral Observatory in Tromsø.

What strange phenomenon could produce such variable characteristics as those of the auroras?

Several observations had alerted physicists to a possible source. As early as 1878, observations of the Montsouris Physical Observatory near Paris showed that the needle of a compass underwent perturbations correlated with sunspots – which we will discuss soon – and with the "great agitations of the Sun" now called eruptions. This link seemed so established that the astronomer Johann Rudolph Wolf (1816–1893), director of the Observatory of Zürich, after having defined an index to calculate the number of sunspots, established formulas to calculate the number of sunspots from the sole examination of the magnetised needle.

A particular event has marked the minds and is, even today, the subject of abundant scientific literature. On the evening of August 28, 1859, auroras were visible as far away as the Caribbean. This specific event will come back several times in the course of this book, so much so that it marks our entire discipline. Abnormal phenomena occurred simultaneously all over the world. In the USA, communications were often made by Morse code, and the country had installed very long transmission cables, interrupted by relays in which the operators received and sent back messages. Spontaneously, some of these relays caught fire. Two English observers, Richard Carrington (1826–1875) and Richard Hodgson (1804–1872), independently observed the Sun using two different pieces of equipment. One projected its image, the other observed (safely) through a telescope. They noted that the Sun was then strewn with sunspots. On September 1, 1859, they observed a lightning flash at the same location, visible to the naked eye. Within a few minutes, magnetic observations at the Kew Observatory, near London, reported strong disturbances of the magnetised needles. Seventeen hours later, the magnetometers recorded an average drop in the Earth's magnetic field of four per cent and the auroras appeared in Cuba, Hawaii, Mexico City…

This is undoubtedly the first widespread known manifestation of space weather on mankind's technology, but it could not be at that time identified as such.

Johannes Jensen (1907–1973) and Maurice Curie (1888–1975), who examined the seasonal variation of auroras in two stations, Saskatoon and Chesterfield, noted with precision a coincidence between the appearance of polar auroras and the presence of important groups of spots on the Sun. In the same vein, Leiv Harang (1902–1970) drew attention to the fact that a strong solar flare occurred on November 25, 1930, while an aurora of exceptional intensity appeared from November 24 to 26. We know today that this was a coincidence, but it was going to help guide researchers on the right track. The Sun seemed to be caught in the act.

Many interpretations were born. As early as 1893, Adam Paulsen (1833–1907), a Danish meteorologist and explorer, launched the idea that the auroras could be produced by rays of charged particles formed in the upper atmosphere under the action of solar radiation. The process would be as follows: the light of the Sun, by a phenomenon still badly understood back then- but which will see its explanation soon after – creates atmospheric electricity which in turn produces auroral radiation. However, this idea did not withstand the observation that auroras appeared in large numbers at night. The idea of charged corpuscles had its attractions. And particularly this one: unlike atoms or neutral molecules, these particles are sensitive to the magnetic field, around which they rotate. This gyration movement could explain the auroral oval proposed since the time of Mayer.

But where do they come from? Could they come from clouds, and rise to 30,000 or 80,000 km before falling back under the effect of gravity? But then, what links them with the Sun? It would be necessary for the rising particles to be electrically neutral. Through interaction with solar radiation, they lose an electron (a process known as ionization), and the resulting ions, now sensitive to terrestrial magnetism, fall back down. This explanation had the merit of establishing a link between solar activity and polar auroras. An experiment seemed extremely convincing. It was proposed by the Norwegian Birkeland (1867–1917) at the end of the 19th century. This young physicist of the aurora had read the articles of 1872 in which Donati suggested that the corpuscles could be emitted directly by the Sun. He then had the idea of a quite extraordinary simulation. He had a glass enclosure built in which he suspended a metallic sphere. Inside the sphere, named Terrella, he installed an electromagnet. In the corner of the box there was an electron gun. Then he created a vacuum, as good as it could possibly be created at the time and which, by chance, corresponded quite well to that of the atmosphere at an altitude of about one hundred kilometres. Shooting the electrons on the sphere, he observed the artificial auroral oval. He had just made the first experimental scientific approach to understanding the mechanism of the polar auroras! In 1896, he presented his experiment as an unexpected effect of the electrons recently discovered by

Joseph John Thomson (1856–1940). Unfortunately, he could not answer questions such as “where are the positive charges? Where is the conductor between the Earth and the Sun in which electrons can move?” Birkeland was ridiculed, misunderstood, and soon morally broken. It was not until well after his death in Tokyo in 1917 that people realised the spectacular progress he had made in geophysics. In his honour, the winds of charged particles that propagate at the pole bear his name, the Birkeland currents: he had the intuition of their origin as early as 1903.

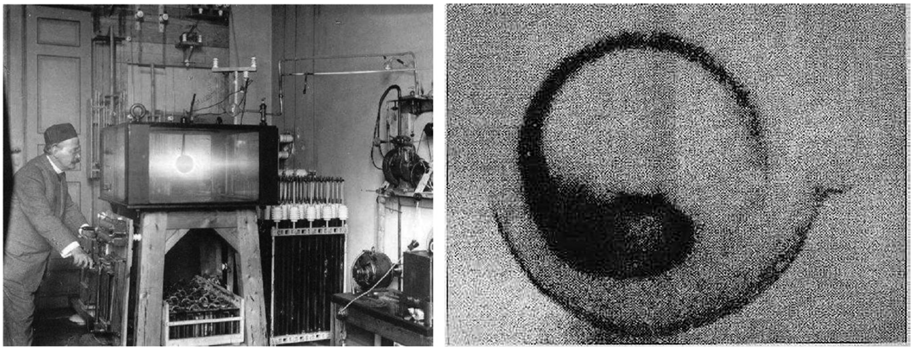


FIG. 1.4 – On the left, Birkeland is in front of his Terrella. On the right, reconstitution of the auroral oval on the Terrella.

Birkeland’s experiment has the great advantage of having received a theoretical explanation in the 1930s. It is to his former student, Carl Størmer (1874–1957) that we owe it. However, this did not prevent other theories from emerging due to the absence of experimental proof through direct measurements in space.

Despite his brilliant intuitions, and perhaps because he did not have the theoretical foundations of the physics of the twentieth century which alone made it possible to explain them, Birkeland was ignored for a long time and even fought. One of his most ardent adversaries was however also one of the most eminent precursors of space weather: Sydney Chapman (1888–1970), of whom we will speak again in chapter two.

The Sun

The Sun provides our solar system with a constant source of energy, and due to its close proximity, shines much brighter in the sky than any other star. The appearance of the Sun in our sky may give the impression that it is unchanging, but in fact, it is quite the opposite. The Sun is a very dynamic star, as we will learn throughout the rest of this book. Its constant emission of energy into our solar system allows for life on Earth: it influences life in important points such as the possibility of photosynthesis for plants, our weather, ocean currents, seasons and even the climate.

It was already in the first century A.D. that Ptolemy (90–168), a very influential Greek mathematician and astronomer, made the first estimate of the Sun-Earth distance: he estimated this distance to be about 1,200 Earth radii (about 7.65 million km), an underestimation of the true distance. This distance was later also found by Copernicus (1473–1543) and then Tycho Brahe (1546–1601). The estimated distance increased to 22 million km with Kepler (1571–1630) and to 110 million with Halley (1656–1742). The Earth resides in an almost circular orbit around the Sun, so the Sun-Earth distance varies slightly with time. Today we know that the average distance between the Sun and the Earth is 149,597,870,700 m (about 23,420 Earth radii). Astronomers call this distance an astronomical unit (1 au).

The Sun emits its energy, which warms the Earth, in a large spectrum, which is the measure of the intensity of light emission as a function of its frequency (or wavelength, also related to its energy). For a long time, observations were limited to observing the Sun in only the visible spectrum; light that we can see with our eyes. However, even from this limited spectrum, we have derived a lot of information. Firstly, all colours are present, from infrared to ultraviolet, with a maximum emission in the green colour. The mixture of all these colours give the Sun its golden hue. Towards the end of the nineteenth century, using the maximum emission in the visible spectrum, physicists deduced the solar surface temperature to be about 6,000 °C, a value that even today remains quite correct. Furthermore, after the first measurements of the visible spectrum, Joseph von Fraunhoffer (1787–1826) noticed that some of the colours were absent. He made the correct interpretation that these colours are being emitted by the Sun but are shortly after absorbed by the solar atmosphere and subsequently cannot be observed as emission anymore. Instead, they are observed as absorption lines. It is only later on, that this absorption observed in the visible spectrum could be linked to specific chemical elements, such as iron, calcium, magnesium, sodium,... Some of the observed absorption lines are connected to chemical elements in Earth's atmosphere, such as oxygen, a link discovered in the nineteenth century during research on auroras (see section *atmosphere and aurora*).

In the 19th century, scientists tried to estimate the total amount of energy the Sun emits, *i.e.*, solar luminosity. The first attempts were made by Claude Pouillet (1790–1868) in 1837 and independently, by John Herschel (1792–1871) during his expedition to the Cape of Good Hope. They constructed a pyrheliometer, which is basically a box filled with water. When it is exposed to sunlight, the water heats up, and measuring the time the water needs to heat up allowed them to estimate the amount of energy that the Sun puts into the water at the Earth's surface. They estimated the solar radiation to be about 1.79 $\mathrm{cal/cm^2/min}$, which was already within 10% of today's accepted value. The difference in measurements comes from atmospheric absorption, which was neglected at that time, making the total solar radiation at Earth's position in the solar system higher than what is effectively observed at Earth's surface. By combining the solar radiation at Earth with the distance from the Sun one finds that the solar luminosity amounts to about 4×10^{26} Watts, or about 4 million times the production of electrical energy by the entire world in 2021.

How can the Sun continuously produce such a large amount of energy? In the absence of knowledge about nuclear fusion at the time, it was difficult for scientists in the 19th century to imagine what processes could produce this absurd amount of energy. The first proposals included the burning of coal or other chemical reactions:

however, this way, the Sun would only maintain its luminosity for a few thousand years. During the first half of the 19th century, geologists had already established that Earth was much older and so the idea was quickly dismissed. It was necessary to find another source of solar energy.

In 1848 the German physicist Julius von Mayer (1814–1878) proposed that the source of solar energy could be the fall of meteorites into the Sun. In this hypothesis, the meteorites would be accelerated by the gravity of the Sun and their kinetic energy would be considerably increased. This energy would then be transferred to the Sun, either directly during the impact or indirectly through friction with the solar atmosphere. At first, this meteorite hypothesis was supported by many notable scientists of the time, such as the Scottish physicist James Waterston (1811–1883) and the great English physicist William Kelvin (1824–1907). However, soon it was realized that the number of meteorites needed to maintain this source of energy would be much higher than the rate of meteorites arriving on Earth. In addition, such a rate of meteorite impacts would change the total mass of the Sun, which in turn would affect the orbit of Mercury.

James Waterston had another idea: he suggested that the Sun could be in contraction under the effect of its own gravity. This contraction leads to a reduction in the gravitational potential energy of the Sun; consequently, this energy would be used to heat up the Sun and it would compensate for the loss of energy by radiation. James Waterson calculated that a reduction of the solar radius by just 5.6 km would generate enough energy to maintain the solar luminosity for 9,000 years! This reduction in radius is almost imperceptible compared to the solar radius (about 0.000008% of one solar radius). This bold hypothesis was based on the theory of the nebula, proposed by Emmanuel Kant (1724–1804) and independently by Pierre-Simon de Laplace (1749–1827). Kant had postulated that a gaseous cloud, a nebula, had collapsed to form the Sun and the solar system. James Waterston understood that this collapse would release gravitational potential energy, energy that, in his hypothesis, would be used to heat the Sun. The German physicist Herman von Helmholtz (1821–1894) adopted this theory of gravitational contraction and calculated that the heat generated by the collapse of a nebula was sufficient to maintain the solar luminosity for 22 million years! These results convinced Kelvin to abandon the meteorite hypothesis. He became such an ardent supporter of the theory of gravitational contraction that it is still known today as the Kelvin–Helmholtz theory. James Waterston, who was not as famous as his two prestigious colleagues, was forgotten… Kelvin and Helmholtz concluded from their work that the Sun, and therefore the Earth, are thus about 20 million years old. This age was, however, not well accepted by the scientific community: biologists claim that the evolution of species would have needed much more time, while geologists estimated the age of certain fossils on Earth at a few hundred million years. At that time however, Kelvin was adamant, based on his Kelvin–Helmholtz theory, that this could not physically be possible and thus geologists had to be wrong.

Then, in 1896, Henri Becquerel (1852–1908), a French physicist, discovered radioactivity in a totally unexpected way: he had left a packet of uranium salts in a drawer, wrapped in a photographic plate. When he later removed the plate, a picture of the uranium salt was impressed within the photographic plates, even though the photographic plates had not been exposed to the Sun! It was as if the uranium was emitting rays of an unknown type by itself. Soon after, in 1907, following the suggestion of the physicist Ernest Rutherford (1871–1937) to use radioactivity as a tool for measuring geologic times, Boltwood (1870–1927), a radiochemist, published a first -rather accurate- list of geologic ages, using radioactivity to determine the ages. The discovery of radioactivity allowed geologists to accurately determine the age of rocks and fossils and by that the age of Earth and of our solar system. They determined that Earth had to be several billion years old, which radically contradicted the age of the Sun assumed by the theory of gravitational contraction. Scientists once again had to look out for another source of solar energy.

While scientists were trying to figure out the source of energy of the Sun, other questions surfaced. Some of these questions concern the variability of the Sun: If the Sun is indeed so old, has it always been as it is today? at what timescales does the Sun change. The first recorded and best-known manifestation of the variability of the Sun is the phenomenon of sunspots. They appear as black spots on the surface of the Sun when observed in the visible spectrum. One of the first recorded observations of sunspots may be found in the Astronomical Treatise of Sòng, a Chinese chronicle spanning the tenth to the thirteenth century, where a "weak and lightless Sun" is mentioned. It still remains a source of wonder how astronomers at the time, long before the invention of the telescope, could see sunspots. Presumably, they looked at the Sun with their naked eye during sunrise/sunset or through the clouds (but please don't risk this yourself! the solar ultraviolet light burns the retina and can make you blind). Nevertheless, the number of sunspots, their shape and their sizes have been on record since January 11, 1077, although, in the beginning, they were only described as "as big as plums". To give an idea of the size, in addition to plums, observers compared them to peaches, glass, hen's eggs and duck's eggs.

In the 17th century, when the telescope was invented, solar observations made a huge step forward. With the help of telescopes, Galileo (1564–1642), as well as Scheiner (1573–1650) and Fabricius (1587–1616), could, independently, clearly see the dark spots that appeared to be on the Sun. However, they were not sure what they were. Are they planets orbiting the Sun, or are they at the solar surface? If they are planets transiting the solar disk, their speed should be constant; however, if they are at the solar surface, they should slowly rotate in, cross the solar disk, and rotate out again. By measuring their speed profile across the solar disk, the second scenario was confirmed, *i.e.*, that the Sun indeed had spots! Sunspots need slightly less than two weeks to cross the visible solar disk, giving a solar rotation rate of about 27 days. From that time on, scientists started to draw the distribution of sunspots on the solar surface, but at irregular intervals. Except for some slow evolution in the sunspots during their passage over the solar disk, they appeared rather boring, and no connection was found to anything else. Then, the interest in sunspots declined again.

It took until 1843 to recover interest, when Heinrich Schwabe (1789–1875), a German pharmacist and amateur astronomer, discovered the solar cycle. In fact, Schwabe's interest was originally not even in sunspots. He was making solar observations in the hope of detecting new planets during their transit across the solar disk, and he had to be persuaded by others to publish his results on

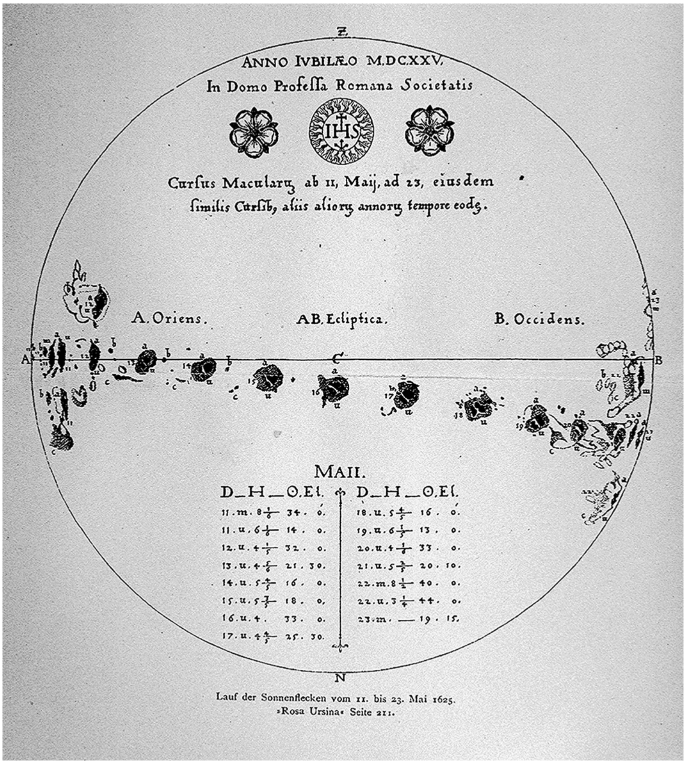
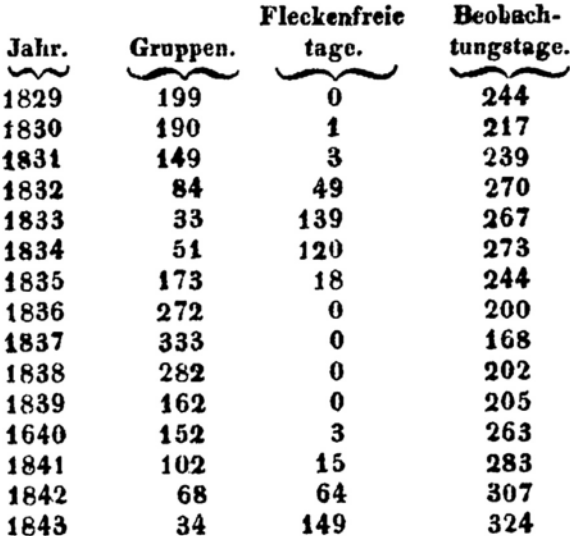


Fig. 1.5 – Scheiner's sunspot drawings as published in Rosa Orsina.

sunspots. His long-term record of sunspots enabled him to see a long-term trend, with years where almost no sunspots were visible on the Sun and years where many were visible. This trend had a period of about 10 years – today, we know that the sunspot cycle has an average length of about 11 years, with 8 years being the shortest and 14 years being the longest recorded. Schwabe was a very patient man to discover the sunspot cycle by daily sunspot observations. This solar cycle is still often referred to as the Schwabe cycle.

The Englishman Richard Carrington (1826–1875) and the German Gustav Spörer (1822–1895) started tracing the appearance of sunspots on the surface of the Sun and remarked that at the beginning of each cycle, they appear about 30° north and south of the solar equator, and as the solar cycle progresses, they slowly appear closer to the equator, while at the same time their total number increases. The figure that appears from tracing the sunspots latitudinal extension with time for more than one full solar cycle resembles the wings of a butterfly and it is still called a *butterfly* diagram today. Astronomers struggled to explain the origin of these observations. The idea that it was linked to the complex rotation of the Sun was quickly discarded, even though this proposition was in fact correct. Instead, astronomers wondered about solar volcanism, about atmospheric vortices, meteorite falls, and the influence of planets such as



Jahr.	Gruppen.	Fleckenfreie tage.	Beobachtungstage.
1829	199	0	244
1830	190	1	217
1831	149	3	239
1832	84	49	270
1833	33	139	267
1834	51	120	273
1835	173	18	244
1836	272	0	200
1837	333	0	168
1838	282	0	202
1839	162	0	205
1640	152	3	263
1841	102	15	283
1842	68	64	307
1843	34	149	324

Fig. 1.6 – Yearly number of sunspot groups (second column), yearly number of days without sunspot groups (third column), and yearly number of observing days (fourth column), as recorded by Heinrich Schwabe.

Jupiter. Some astronomers went as far as to assume that the spots were fictitious – but of course, they were not. The mystery about sunspots seemed to thicken.

Besides sunspots, other solar phenomena intrigued the early solar physicists. During eclipses of the Sun by the Moon, the solar limb often seemed to be the seat of enormous arches of fire, called prominences. The first prominences were identified in 1239, but back then, they were considered to be optical illusions. In 1842, during a solar eclipse, the observing conditions in Europe were particularly favourable and several large prominences were visible beyond the solar limb. The Englishman George Biddell Airy (1801–1892), who observed this eclipse in Italy using a 1.4-inch refracting telescope, made the first detailed description of this region. Were these prominences of volcanic origin? Were they mountains in movement? Some of them appeared to rise to altitudes of 300,000 km in just a few minutes. Since the prominences looked like jagged mountains, he decided to name these regions observed off the solar limb, the *Sierra*, *i.e.*, the mountain chain. The French François Arago (1786–1853) and Frédéric Petit (1810–1865) estimated the diameter of these prominences to be about 80,000 km, six times the Earth's diameter, thus excluding prominences from being optical illusions. In 1868, Joseph Norman Lockyer (1836–1920) gave the *Sierra* its present name: the *chromosphere*, the coloured sphere.

A few decades later, in 1868, Jules Janssen (1824–1907) made a long journey to India to watch another solar eclipse. He aimed to study the spectrum of the chromosphere during this eclipse, and, indeed, he found several emission lines. Between the two known sodium lines in the yellow part of the solar spectrum, he observed a

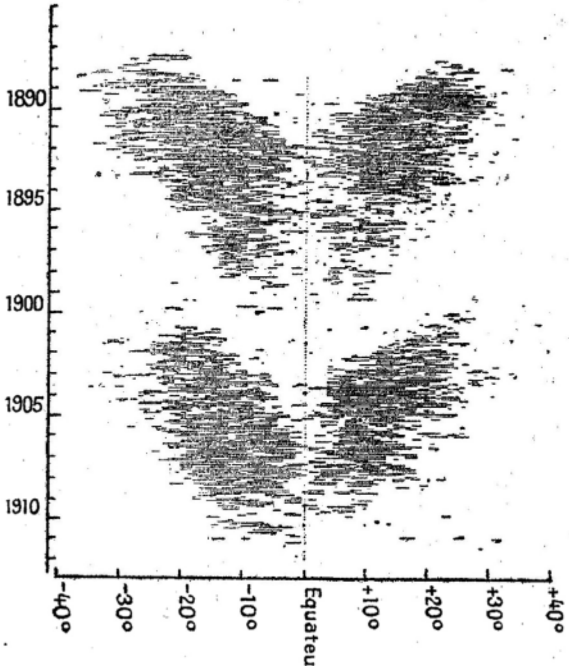


Fig. 1.7 – The first butterfly diagram. The ordinate shows the years, starting from 1886 and the abscissa shows the solar latitude. Each point represents a spot or a group of sunspots. Observers try to show on the drawing the apparent extent that they have on the Sun (after Couderc, 1932).

very bright, but unknown spectral line. What could this line be, as it was not listed as any of the known elements? As the line was extremely strong, he tried to observe the same line again the next day after the eclipse had passed by looking right next to the solar disc. It turned out that this line was visible even without an eclipse blocking the Sun's light! How can this unknown line be explained? One considered possibility was that the Sun contained a new element that was not yet detected on Earth; Norman Lockyer and Edward Frankland (1825–1899) called this new and at this time hypothetical element helium, from the Greek *helios*, Sun. Helium was later isolated in the laboratory in 1895, confirming the hypothesis. It was an extraordinary discovery: an element was revealed for the first time outside the Earth.

Inspired by Janssen's observations of the chromosphere, George Hale (1868–1938), Henri Deslandres (1853–1948) and John Evershed (1864–1956), invented and perfected the spectroheliograph during the decades of 1880 and 1890. The spectroheliograph allowed them to observe a single wavelength in the solar spectrum, and by that to observe the chromosphere at the solar disk and offlimb at different temperatures and heights. Due to their invention, it was realised that prominences off the solar limb, observed during eclipses, were related to the phenomenal explosions

Fig. 1.8 – Picture of the Jules Janssen Expedition.

close to the solar surface. The more sunspots were observed on the solar surface, the more numerous these eruptions were. Astronomers now believed that the Sun was not a quiet star in the sky, but that the Sun was indeed active.

Now that scientists realised the Sun was a very active star, and not at all static, they started wondering whether the Sun could affect our Earth, even though they are more than 150 million km apart. In 1838, Sir Edward Sabine (1788–1883) consulted with Alexander von Humboldt (1769–1859) in Berlin. They asked the Royal Society to establish records of magnetic observatories throughout the whole world and superintended the effort within the British empire. Three magnetic observatories were founded: one in St Helena, located in the South Atlantic Ocean, one at the Cape of Good Hope, and one in Toronto, Canada. Sir Sabine collaborated with observatories at the Great Bear Lake and Barrow Strait in Canada and with Melville Island in Australia. Having access to a long-time series of magnetic field measurements of Earth's magnetic field, he discovered that geomagnetic disturbances at Earth manifested a regular diurnal variation and an irregular long-term cycle. In 1852, he stated that the Sun's irregular 11-year sunspot cycle was absolutely identical to the Earth's irregular 11-year cycle of geomagnetic disturbances. As a result, the Sun must be influencing Earth.

Over the next decades, more evidence of the Sun's influence was discovered. Edward Maunder (1851–1928), an English astronomer, compared the timestamps of 19 great geomagnetic storms and found that, at each of them, large sunspot groups had been observed close to the centre of the solar disk. He further identified a group of geomagnetic storms that appeared at a 27-day interval, which seemed to match the recurrence interval of a sunspot group, *i.e.*, the solar rotation rate of about

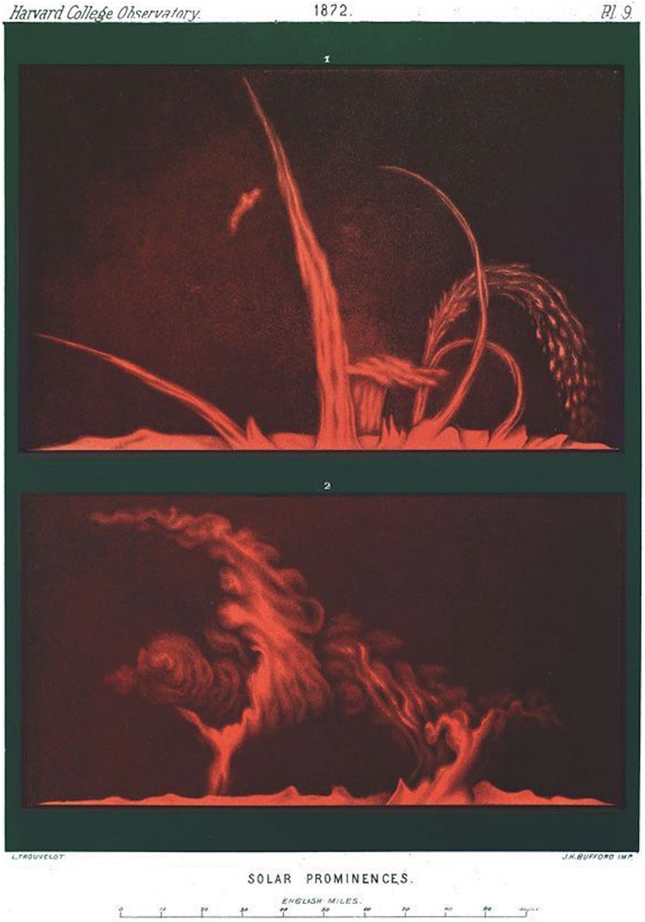


Fig. 1.9 – Prominence drawing published in 1872 in the Annals of Harvard College by Trouvelot, a later staff member of Jules Janssen.

27 days. Emile Marchand (1852–1914), then a meteorologist at the Lyon Observatory, became interested in the problem of magnetic storms. He carefully compared the magnetic disturbances measured by the Lyon Observatory with his observations of the Sun. Marchand noted in 1887: "Magnetic disturbances occur when a region of activity of the Sun passes the central meridian". What he calls an active region is a region that contains sunspots but also bright ones called faculae. Marchand concludes in his 1904 publication: "(...) during magnetic storms, the disturbing currents of the Earth's field are located, at least partially, in the upper regions of the atmosphere". Building on these discoveries, in 1929, William Greaves (1897–1955) and Harold Newton (1893–1985) found that larger storms appeared as a one-time occurrence, while smaller geomagnetic storms appeared more frequently one after another with a 27-day interval. However, not all of the smaller storms could be

associated with a group of sunspots. Therefore, Bartels (1899–1964) suggested in 1932 that there could be mysterious regions, which he dubbed M-regions, on the Sun that are fixed at the solar surface. These M-regions could be but do not necessarily have to be related to sunspots. So, was the Sun responsible for the geomagnetic disturbances, or was it just a pure coincidence? Not everyone was convinced about the relationship between the Sun and the Earth.

One of the greatest adversaries of a solar-terrestrial relationship was Lord Kelvin – who, at this time he was already a renowned scientist. He came to the following conclusion regarding the Sun being the source geomagnetic storms [our notes added within brackets]:

> "To produce such changes as these [of the Earth's magnetic field during geomagnetic storms] by any possible dynamical action within the Sun, or in his atmosphere, the agent must have worked at something like 160 million, million, million, million horsepower [12 × 10^35 ergs/s], which is about 364 times the total horsepower [3.3 × 10^33 ergs/s] of the solar radiation. Thus, in this eight hours of a not very severe magnetic storm, as much work must have been done by the Sun in sending magnetic waves out in all directions through space as he actually does in four months of his regular heat and light. This result, it seems to me, is absolutely conclusive against the supposition that terrestrial magnetic storms are due to magnetic action of the Sun, or to any kind of dynamical action taking place within the Sun, or in connection with hurricanes in his atmosphere, or anywhere near the Sun outside. It seems as if we may also be forced to conclude that the supposed connection between magnetic storms and sunspots is unreal, and that the seeming agreement between the periods has been a mere coincidence." (Kelvin, 1892)

His strong opposition to a solar-terrestrial relationship slowed down the scientific progress of the understanding of this relationship – but fortunately, he could not halt it.

The First Steps

Two separate paths were paved by scientists. One path explored Earth's magnetism, investigated beautiful auroras and discovered details about Earth's atmosphere. The other explored the spectrum of the Sun, speculated about its energy source, learned about Earth's age and consequently about the Sun's age and uncovered the fascinating variability of the Sun by mapping the solar cycle. At the time of the Carrington event of 1859, an event that is believed to be the strongest space weather event recorded by mankind, these two journeys did not realise yet that soon their paths would undoubtedly intertwine. Many knowledge gaps were still present, knowledge about the Sun and the space environment of the Earth was still fragmentary, the understanding of solar activity was sometimes based on assumptions and the connection between the Sun and Earth was only considered by just a few observers. However, this would soon change. In the span of only one century, during the space age, which will be explored in the next two chapters, scientific revolutions will reveal the interaction between the Sun and the Earth so that the two paths will finally merge to create space weather and space climate.

References

Couderc Paul, Dans le champ solaire, Gauthier-Villars Editeurs, 1932.

Flammarion Camille, Popular astronomy: a general description of the heavens, New York, D. Appleton, 1890 (accessible in the Interntet Archive).
Original in French: L'astronomie populaire, Marpon et Flammarion éditeurs, 1882.

le Gars Stéphane, Dortous de Mairan et la théorie des aurores polaires : trajectoire et circulation d'une idée de 1733 à 1933, Armand Colin ed., « Revue d'histoire des sciences » 2015/2 Tome 68,pages 311 à 333, ISSN 0151-4105 ISBN 9782200930141, https://www.cairn.info/revue-d-histoiredes-sciences-2015-2-page-311.htm.

Génot Vincent. Étude des phénomènes d'accélération de particules dans les régions aurorales des magnétosphères. Astrophysique [astro-ph]. Université de Versailles-Saint Quentin en Yveline, 1999.

Harvard College Observatory, 1876. Annals of the Astronomical Observatory of Harvard College v8. Publisher: Wilson, J., and Son.

Hayakawa Hisashi, Fujii Yuri, Murata Koji, Mitsuma Yasuyuki, Cheng Yongchao, Nogami Nagatoshi, Ichikawa Kohei, Sano Hidetoshi, Tsumura Kohji, Kawamoto Yukiko, Nishin Masaki, Three Case Reports on the Cometary Plasma Tail in the Historical Documents, JSWSC, Section Agora, 2020.

Hayakawa Hisashi, Mitsuma Yasuyuki, Ebihara Yusuke, and Miyake Fusa, (2019) The Earliest Candidates of Auroral Observations in Assyrian Astrological Reports: Insights on Solar Activity around 660 BCE, *The Astrophysical Journal Letters* **884**, L18, 2019, https://doi.org/10.3847/2041-8213/ab42e4.

Kelvin, The Lord (1892) The anniversary of the royal society, *Nat.* **47**, 106–111.

Malin S.R.C., Barraclough, D.R. (1991) Humboldt and the earth's magnetic field, *Quart. J. Royal Astron. Soc.* **32**, 279.

de Mairan Dortous, Lettre de M. De Mairan, secrétaire perpétuel de l'Académie Royale des Sciences à madame la Marquise du Chatellet, février 1741, https://gallica.bnf.fr/ark:/12148/bpt6k3586v.texteImage.

Rosmorduc Jean, Une histoire de la physique et de la chimie, collection Point Sciences, Ed. du Seuil, ISBN 2 02 008990-4, 1985.

Schwabe H., Sonnen-Beobachtungen im Jahre 1843. In: Some Aspects of the Earlier History of Solar-Terrestrial Physics (2004). Editor Schröder, Wilfried.

Ursina Rosa. Bracciani [Bracciano]: apud Andream Phaeum Typographum Ducalem, Impressio coepta anno 1626. finita vero 1630. ETH-Bibliothek Zürich, Rar 10152, https://doi.org/10.3931/e-rara-556 / Public Domain Mark.

Chapter 2

The Time of Discoveries

In the first chapter we witnessed a slow maturation of humanity on the understanding of the Earth, the Sun, and its close space environment. Some of the observations still resisted explanation, but everything seemed to have an answer that follows the laws of physics sooner or later. Nevertheless, scientists did not imagine that celestial bodies were in interaction with each other, other than gravitationally and that their interactions would soon be at the origin of an increasing complexity in our vision of the space environment.

The Sun's Energy Problem

Let us first return to the problem of the energy source of the Sun. We have seen that the theory of the Sun's energy supply by gravitational contraction is not compatible with the age of the fossils on Earth as derived by geologists. Gravitational contraction could justify the energy necessary for the Sun to shine for only about 20 million years. Nineteenth-century physics had no solution to this problem: none of the known energy source could explain the Sun's brightness over billions of years. But everything changed with the discovery of radioactivity. After this discovery by Henri Becquerel (1852–1908), Marie Curie (1867–1934) realised that certain rocks released extraordinary amounts of energy. She called this phenomenon radioactivity. However, the origin of radioactivity remained unknown.

Atoms were supposed to be indestructible: the word atom comes from the Greek *atomon*, indivisible. However, the New Zealander Ernest Rutherford (1871–1937) and the Englishman Frederick Soddy (1877–1956) realised in 1903 that radioactive elements transform into other elements. During this transformation, these elements release a prodigious amount of energy, far more than any chemical reaction. It was an extraordinary, almost unimaginable result: pure alchemy! In 1908, Rutherford made a discovery that was perhaps even more extraordinary: the atomic nucleus. Until then, it was thought that the mass of atoms was uniformly distributed throughout the volume of the atoms. Rutherford showed that almost all the mass of atoms is located in a nucleus that is much smaller than the atom itself. The rest of the mass is distributed among the electrons that circulate around the nucleus. And Rutherford understood that radioactivity was a phenomenon that concerned only the nucleus.

DOI: 10.1051/978-2-7598-3610-9.c002

Physics was to undergo several revolutions at the beginning of the 20th century. In 1905, Einstein (1879–1955) published the theory of special relativity. Among the predictions of this theory there is the famous equation $E = mc^2$. Mass can be converted into energy, lots of energy! If we find a way to convert 1,000 kg of matter completely into energy, we will get 9×10^{19} J of energy. This energy can meet the energy needs of the entire European Union for 1 year. And the physicists who were still looking for the source of solar energy did not miss this information.

In 1915, the American William Draper Harkins (1873–1951) proposed the fusion of hydrogen into helium as the missing source of energy. Indeed, four hydrogen nuclei (protons) can fuse to form a helium nucleus. However, if we carefully measure the masses of the hydrogen and helium nuclei, we find that the helium nucleus is slightly lighter than the 4 hydrogen nuclei: a difference of 0.7%. This missing mass has to be transformed into energy. It is a small difference in mass, but thanks to Einstein's equation we understand that this small difference can provide a lot of energy. Enough energy, in fact, to keep the Sun bright for 9 billion years. At last, a source of energy powerful enough to solve the problem of solar energy was found. The exact reaction was determined in 1939 by Hans Bethe (1906–2005). The previous year, the 4th Washington Conference on Theoretical Physics was dedicated to the problem of the energy of stars. Hans Bethe was there and, on the train back to Cornell University – where he was a professor – he determined all the important nuclear reactions that contribute to the fusion of hydrogen into helium.

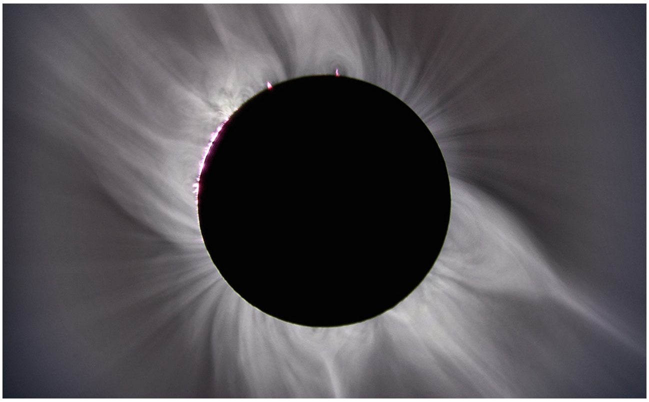


Fig. 2.1 – The annular eclipse of 29 March 2006 taken from Sidi Barrani, a location in Egypt near the eastern border of Libya. The solar corona is very bright, with several coronal loops. Its shape, wide at mid-latitudes and squashed at the poles, is indicative of moderate solar activity. The chromosphere appears as a red border on the edges of the Moon that hides the Sun, and several prominences are revealed (Credit: Jean Mouette/IAP-CNRS-SU).

The Solar Corona

The solar corona is a clearly visible luminous halo that sits on top of the chromosphere and surrounds the Sun during a total eclipse. The corona has probably been observed for as long as mankind has existed. But the first clear description probably dates back to the year 968 with the observation of an eclipse by Byzantine historian Leo Diaconus (ca. 950–994): *"... at the fourth hour of the day ... darkness covered the Earth and all the brightest stars shone forth. And is was possible to see the disk of the Sun, dull and unlit, and a dim and feeble glow like a narrow band shining in a circle around the edge of the disk"*.

Until the 19th century, it was difficult to decide whether the corona was part of the atmosphere of the Moon or that of the Sun. The answer came from a new technology: photograph. Shortly after its invention, Louis Daguerre (1787–1851), the inventor of the daguerreotype process, tried to photograph the Moon in 1839. The long exposure time required by this technique, coupled with imperfect tracking of the Moon, meant that the final result was only a blur! But they learned from this failure, and John William Draper (1811–1882) managed to take the first unblurred photo of the Moon the following year. The Sun was the next target: Léon Foucault (1819–1868) and Hippolyte Fizeau (1819–1896) succeeded to image the solar disk in 1845. On July 28, 1851, Johann Julius Friedrich Berkowski finally managed to take the first photo of the solar corona during an eclipse, having an exposure time of 84 s. On 18 July 1860, the Italian priest and astronomer Angelo Secchi (1818–1878) and the Englishman Warren de la Rue (1815–1889) were in Spain to observe and photograph a total eclipse of the Sun. They chose two different observation sites, 500 km apart. If the coronal features are part of the moon, the two photos observed from slightly different viewing angles should look slightly different (this is call the parallax effect); but if they are features of the Sun, both photos should be identical. By comparing the features of the corona, Secchi and de la Rue showed that the corona was part of the Sun.

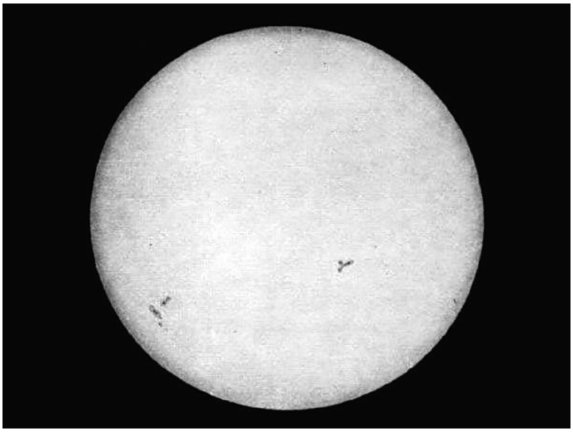


FIG. 2.2 – First-ever photography of the Sun by Foucault and Fizeau, 1845.

During an eclipse in 1869, Charles Augustus Young (1834–1908) and William Harkness (1837–1903) used the newly invented spectroheliograph to measure the spectrum of the solar corona and found an unknown green emission line. They proposed that this line, analogous to the helium line observed in the chromosphere, resulted from another unknown element. This hypothetical element was called coronium. But, unlike helium, coronium could not be found in the laboratory.

A breakthrough needed to wait for a new instrument able to observe the corona outside eclipses: the coronagraph. In 1931, the coronagraph was invented by the Frenchman Bernard Lyot (1897–1952). It is an ingenious optical device that makes it possible to occult the dazzling solar disc and reveal the solar environment: the prominences and the corona. The coronagraph is basically a refractor telescope. Lyot placed an occulting disc, the so-called Lyot mask, in the focus of the telescope where the image is formed to cover exactly the image of the solar disc. In this way, the photosphere is hidden, allowing the outer part of the image, the corona, to be seen. This is an artificial eclipse! But hiding the solar disc is not enough, others have tried before Lyot without ever achieving a satisfactory result. Indeed, the occulting disc does not block the sunlight diffracted by the lens, as well as the light diffused by the humidity, dust and pollution of the sky. This stray light is brighter than the corona and prevents from imaging the latter. To solve the problem of light diffracted and scattered by the lens, Lyot had the brilliant idea of introducing a diaphragm stop into the instrument, (today known as the Lyot stop). To solve the problem of the light scattered by the sky, Lyot chose an observation site high in the mountains where the air is as pure as possible: the Pic du Midi in the Pyrenees.

FIG. 2.3 – Bernard Lyot at Pic du Midi.

The coronagraph allowed him and other scientists to study the evolution of the corona even outside of eclipse time; but, nevertheless, the coronium line stayed a mystery.

Then, in 1939, the German Walter Grotian (1890–1954) saw an article by the Swedish spectroscopy specialist Bengt Edlén (1906–1993) on the forbidden transition lines of highly ionised atoms in supernova. He realised that one of the unidentified coronal lines matched one of the lines identified by Edlén in supernovae studies, and suggested that the coronal spectral lines could be of highly ionised atoms. But then, the corona must be extremely hot, much hotter than the photosphere. Bengt Edlén read that article and realised at once that Grotian was right. In the following years, he dedicated himself to this subject, published a preliminary report in 1941, and a full article that explained the mysterious coronal line in all detail in 1943. The mysterious coronal line was not a new element, but something equally extraordinary: iron, ionised 13 times. This means that 13 of the 26 electrons of the iron atom have been torn off! It is no wonder that astronomers have not been able to identify it before. The process of producing a 13-ionised iron is very difficult as it is a very energy-intensive process. If iron is highly ionised in the corona, the conclusion is inevitable: the temperature of the corona is at least 1 million degrees Celsius.

Also in 1941, the Swede Hannes Alfvén (1908–1995) heard about the idea of a million-degree hot corona by Grotian and Edlén and tackled this idea from a further viewpoint. He derived six theoretical reasons, based on observations, why the corona should be a million degrees hot. The simplest of the six reasons is that to maintain the observed off-limb extension of the gaseous corona under the effect of solar gravitation, the gas would need to have a temperature of about one million degrees. Already at this time, he derived the density of the corona to be 10^8 particles per cm^3 and a required heating rate of about 200 $\mathrm{W/m}^2$ (which is not too far from today's assumed value of about 400 $\mathrm{W/m}^2$). Furthermore, he derived that particles in the solar corona have to propagate along the curvature of the magnetic field in the solar corona, and thus structures in the solar corona outline its magnetic field topology.

This high temperature of the solar corona has two direct consequences. First, the wavelength of the light emitted by an object depends on its temperature; the higher the temperature, the shorter the wavelengths emitted. At the typical temperatures found on Earth, thermal emission is almost entirely in the infrared. As the temperature increases, the emission peak shifts to red: this is why iron turns red when heated. At about 30,000 °C, almost all the energy is radiated in the ultraviolet range. At the temperature of the solar corona, the emission is mostly in the extreme ultraviolet (the most energetic ultraviolet) and soft X-rays (the least energetic X-rays). Therefore, the corona must be very bright in the extreme ultraviolet, much brighter than in the visible part of the spectrum!

Second, the question arises on how the corona can have a higher temperature than the photosphere. The corona lies above the photosphere, which has a temperature of just about 6,000 °C. One would expect that the temperature decreases farther away from the Sun. Where does the energy that heats the corona come from? This mystery has not been fully solved even today.

Emergence of New Physics and a New Concept: Plasma and the Solar Wind

Sydney Chapman, who was already mentioned, proposes that electron beams (cathode rays as they were then called) from the Sun were responsible for the magnetic disturbances on the ground. This was an interesting hypothesis, but how could such a beam be collimated over the Sun-Earth distance? This would require colossal electric fields, something which is hardly possible.

Birkeland was convinced that space is filled with electrons and protons. As early as 1908 he wrote in his notes about an auroral expedition in the winter of 1902–1903:

> "It seems to be a natural consequence of our views to suppose that all space is filled with flying electrons and electric ions of all kinds."

He then proposed a bold idea in a 1916 paper:

> "From the physical point of view, it is very probable that these new solar rays are exclusively neither positive nor negative rays, but are rays of both kinds."

Note that he cautiously took up the dominant idea of the time, namely that the Sun and the Earth are bathed in an empty space and that the Sun sporadically ejects jets of electric particles. However, we have seen that he believed – correctly – that space was continuously filled with these solar particles.

Frederick Lindemann (1886–1957) proposed the same idea three years later (independently, as it seems): *"the Sun emits a soup of electrically charged but globally neutral particles, i.e. with as many positive as negative electric charges"*. This proposal was doubly attractive: it combines the presence of electric charges, which are suspected of being the cause of the aurora, and the electrical neutrality of the whole, which dispenses electric fields in space to maintain particle beams. The concept of plasma was born. The term plasma was introduced in 1928 by Irving Langmuir (1881–1957).

Carrington, Birkeland, and others suggested in the past that the Sun could expel electrically charged particles that affect the Earth's atmosphere and magnetic field in order to explain the geomagnetic disturbances at Earth. At the beginning of the 20th century, this idea gained ground and started to get accepted more and more. The movement of charged particles in a magnetic field was already well described, particularly thanks to the work of James Maxwell. What remains to be understood was the nature of these particles, how the Sun expels them, and how they were transported from the Sun to the Earth.

The concepts and the calculations from the early work on the Sun's corona, as well as the correlation between sunspot numbers and the frequency of the occurrences of geomagnetic storms, indicated that the interplanetary space around the Sun is bathed in a flow of particles that affects the Earth. However, there was a lack of evidence. The first strong hint came from comets. In 1908, Arthur Eddington (1882–1944) studied the shape of the comet *Morehouse* and found that there must be a force opposite to the Sun's attraction that determines its ovoid shape followed by the tail. He surmised that it could be due to *"a swarm of ions coming from the Sun"*.

In the same vein and without any obvious link with Eddington's work, the German Cuno Hoffmeister (1892–1968) also noted in 1943 that comets have their tails constantly oriented away from the direction of the Sun, as if they were being pushed by some force. His compatriot Ludwig Biermann (1907–1986) came to the same conclusion and spoke in 1951 of 'solar corpuscular radiation'.

The famous Halley's comet has been observed for centuries. As early as 760, the Chinese chronicles of Zuqnin testify that two tails follow the comet, one "narrow and dusty", the other "wide and darker". The same observations are made in 837 on a new passage of Halley. In 1577, two separate observations, in Belgium and Syria, testify the same phenomenon on the comet inelegantly named "C/1577 V1". We now know that the large dust tail is due to the mass loss of the comet and that it follows its trajectory. The other tail is of a more subtle origin, the physics of which will be explained in the following chapters. It has long been a subject of perplexity. On February 27, 1843, a comet passed close to the Sun in such a configuration that its tail could be followed for several hours. Surprise! It is always opposite to the Sun. Astronomers of the time saw proof that there is a continuous medium in the universe, the ether, which undulates under the influence of the comet itself.

We now know that the mechanism occurs in three stages. First, as it approaches the Sun, the icy matter of the comet sublimates: the surface vaporises. Then, the extreme ultraviolet radiation tears electrons from the molecules of this gas, which immediately becomes sensitive to the electromagnetic environment and is carried away in the solar wind. The solar wind? Yes, comets prove that the Sun permanently emits wind, which we will discuss again and again, as it is at the origin of space weather.

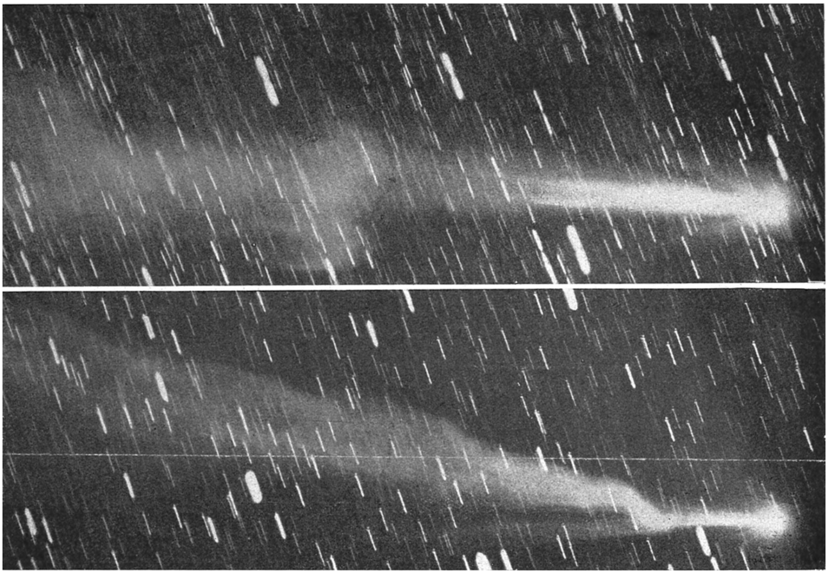


Fig. 2.4 – Comet Morehouse, photographed at the Lick Observatory on 15 and 16 October 1908 (Credit: F. Quenisset, from Dans le champ solaire, Paul Couderc, Ed. Gauthier-Villars, 1932).

A full theory on this ‘solar corpuscular radiation’ was still missing. The theoretical work started in 1939 when Hannes Alfvén proposed the theory of magnetised fluids, magnetohydrodynamics (MHD). In general, MHD takes the equations of classical fluid mechanics, the Navier–Stokes equations, and couples them to the equations of electromagnetism, Maxwell equations. MHD is satisfactory for describing weakly colliding fluids, *i.e.* where the collision frequency between constituents is low compared to other characteristic frequencies (such as the gyration frequency). In the ideal MHD, a simplification of the MHD, we consider a high magnetic Reynolds number. Under this assumption, the plasma is a perfect conductor. One of the important results of ideal MHD is Alfvén’s theorem: the magnetic flux through a plasma element is conserved during the flow. In plain English, this means that if a plasma element is traversed by magnetic field lines, it will remain on those field lines. Conversely, if a plasma element is not bound to magnetic field lines from its genesis, it cannot become so later. This phenomenon is referred to as *freezing* of the magnetic field in the plasma. The implications for the Sun-Earth relationship are immense: as the solar wind escapes from the Sun, it carries with it the solar magnetic field, which we will call the interplanetary magnetic field. As the emitted solar wind approaches the geomagnetic field, it cannot penetrate it. Finally, Alfvén’s approach made it possible to describe the waves that propagate in magnetised plasmas, including the so-called Alfvén waves, which can be seen as oscillations of the magnetic field.

Hannes Alfvén

Hannes Alfvén received his doctorate from the University of Uppsala in Sweden in 1934 for his work on high-frequency electromagnetic waves, after which he taught physics there. In 1940 he became a professor of electromagnetism at the Royal Institute of Technology (KTH) in Stockholm. His seminal paper in which he presented his theory of magnetohydrodynamics was submitted to the journal *Terrestrial Magnetism and Atmospheric Electricity* but was rejected. Alfvén was initially obliged to publish his results in Swedish and German journals with limited circulation. Alfvén raised thus the problem of scientific journals’ reluctance to accept new ideas and criticised the peer review system:

> “When I describe the phenomena according to this formalism most referees do not understand what I say and turn down my papers. With the referee system which rules US science today, this means that my papers rarely are accepted by the leading US journals. Europe, including the Soviet Union, and Japan are more tolerant of dissidents”.

The MHD theory he was developing came to light at symposia and in a somewhat roundabout way in a 1942 paper on waves propagating in magnetised media.

Hannes Alfvén was awarded the Nobel Prize in Physics in 1970 for his work, but the image of a researcher with crazy ideas stuck to him and he was never particularly well regarded in his time. This reputation came also from his opposition to the expanding Universe and the Big Bang model, and he, together with Oskar Klein (1894–1977), developed an alternative model.

Fig. 2.5 – Hannes Alfvén in 1942.

It was not until the end of the 1950s that the term solar wind, and the explanation that goes with it, was proposed by another great man of solar physics: Eugene Parker[1] (1927–2022). Parker was interested in the solar corona. He first calculated the thermal energy required by the particles in the corona to prevent them from falling back under the effect of gravity. He found a temperature of the order of a million degrees Celsius, consistent with the discovery of the 13 times ionised iron in the solar corona. Considering this temperature, he tried to explain the corona in hydrostatic equilibrium, *i.e.* a corona in which the pressure of the gas is precisely counterbalanced by the gravitational force produced by the solar mass. He soon realised that this explanation was not satisfactory because it would imply that the pressure at infinity had to be much greater than what he considered to be a typical value for the pressure of the interstellar medium. He then considered another hypothesis, that of a hydrodynamically expanding corona: solar gas continuously escapes radially in all directions, follows the pressure gradients, and thereby becomes accelerated to supersonic speeds propagating radially away from the Sun. Since the plasma follows the pressure gradients analogous to wind streams that follow pressure gradients at Earth, he named these outgoing plasma flows the *solar wind*. In the same paper, he described the evolution of the solar magnetic field in interplanetary space. Since, according to Alfvén's theory of ideal MHD, the magnetic field is frozen in the outflowing plasma from the Sun to interplanetary space,

---

[1] Since Eugene Parker died in 2022, the reader will find many biographies of him on the web. The authors suggest starting with the NASA one, since Eugene Parker worked so much with this agency: https://www.nasa.gov/press-release/nasa-mourns-passing-of-visionary-heliophysicist-eugene-parker.

and since the Sun is rotating, the Sun's magnetic field lines must wind up like a spiral in the plane of the ecliptic. We called this spiral the Parker spiral.

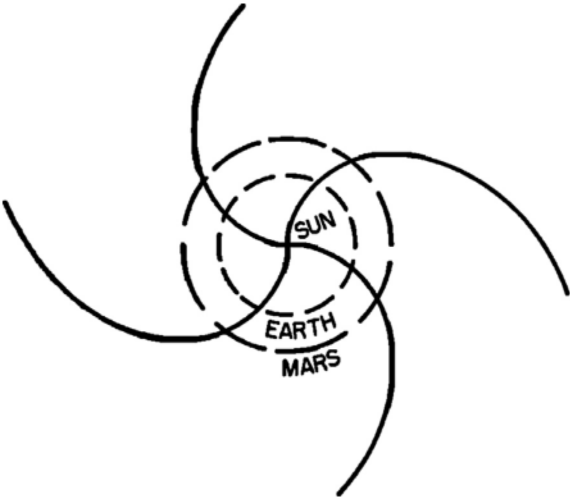


FIG. 2.6 – Projection in the ecliptic plane showing four lines of the solar magnetic field that wrap around as the solar wind propagates radially as the Sun rotates (counter-clockwise in this diagram). The orbits of the Earth and Mars are also shown (Figure taken from a 1958 paper by Eugene Parker). The concept can be understood by comparison with a spraying water hose, creating a spiral structure when rotating.

Discovery of an Electrically Conductive Atmospheric Layer: The Ionosphere

In parallel to the first ideas on the solar corona and the solar wind, the understanding of the structure of Earth's atmosphere advanced, surprisingly finding plasma within it as well. Measurements made on the ground during the 18th and 19th centuries revealed spatial and temporal variations in the intensity and orientation of the geomagnetic field. Spatial variations mean that the intensity and orientation of the magnetic field at ground level differ from one place to another at a given time, while temporal variations refer to variations in the same quantities over time at a given location. Spatial variations during geomagnetically quiet conditions are due to the fact that the Earth's internal magnetic field and the local magnetization of the rock (lithospheric field) are not uniform. These spatial variations originate from the planet and evolve only in the long term; we will not consider them here. The temporal variations, on the other hand, present periodicities that were quickly linked to external sources.

Fig. 2.7 – A very moving photograph: in March 2017, Eugene Parker (left), professor emeritus at the University of Chicago, visits the space probe that bears his name and was launched on August 12, 2018: the NASA Parker Solar Probe. He stands here on the launch pad with Dr Nicky Fox (right), with the Delta Heavy in the background ready to launch Credit: NASA/Johns Hopkins APL/Ed Whitman.

Thanks to Lorentz (1853–1928), 19th-century researchers were aware of the phenomenon of magnetic induction and therefore attributed the magnetic variations measured on the ground to the propagation of electric currents. But where do these currents propagate? In the ground? In the atmosphere?

The discovery of the atmospheric conductive layer is closely linked to the development of radio frequency technologies, led by Nikola Tesla (1856–1943) and Guglielmo Marconi (1874–1937). Marconi, who had already carried out conclusive tests of radio links over short distances, carried out the first transatlantic radio link experiment on December 12, 1901 between a transmitter placed in Poldhu in Cornwall, England and a receiver in Saint John's, Newfoundland, Canada. For three days, the operators in Poldhu were instructed to transmit the Morse code of the letter S (3 points) continuously at 6 pm. At the time, critics scoffed at this challenge: given the Earth's curvature and the rectilinear nature of radio wave propagation, the signal reception was deemed impossible. However, Marconi successfully received the signal at Newfoundland, using as antenna a simple copper cable carried 135 m high by a kite. Nevertheless, this link over the horizon could only be made after at least one reflection of the emitted wave from a reflective atmospheric layer as envisaged by Kennelly (1861–1939) and Heaviside (1850–1925) in 1902, later called the Kennelly–Heaviside layer to acknowledge them.

The Marconi Experiment

Let us stop to discuss the Marconi experiment. Even if it really marks an important step in the development of radio communications and research on the upper atmosphere, and even if it was worth the Nobel Prize in Physics to its author in 1909, it is still shrouded in mystery and doubt. Indeed, there are reasons to think that Marconi failed and did not hear the Morse code signal that was sent to him. Why questioning this experiment? Already at that time, there was doubt. With the equipment he had, he could not control or measure the frequency of the emitted wave, but we can suppose that it emitted at a frequency of a few hundred kHz. In this frequency range, the maximum signal range is about 500 km during the day, longer at night when the reflective ionospheric layer is higher and more favourable to long-distance transmission. The distance between the transmitter and Marconi's receiver was about 3,400 km. Moreover, during the three days of the experiment, Marconi declared that he heard the three points of the S only once or twice! And no one except him witnessed it. So what did he hear, if anything at all? Maybe sizzling sounds or the radio disturbances associated with lightning? Experiments attempted in the 2000s to reproduce the experiment with modern or old equipment ended in vain.

Fig. 2.8 – Guglielmo Marconi and others set up the kite carrying the antenna for the first transatlantic wireless communication, in St. John's, Newfoundland, on December 12, 1901.

In 1924, Edward Appleton (1892–1965) imagined and set up an experiment that led to the measurement of the altitude of this reflective layer. The idea was to send a radio signal from a transmitter to a receiver, both on the ground. A part of the emitted signal will reach the receiver directly while another part will reach it *via* an atmospheric reflection at a certain altitude. In principle, the difference in propagation time gives double the altitude of the reflecting layer. In practice, Appleton relied on the phenomenon of interference: if two waves of the same frequency are superimposed and their paths differ by an integer number of wavelengths (the length of a pattern of the wave), then we obtain an increase in the total signal and call it constructive interference. Conversely, if the paths differ by an odd number of half-wavelengths, the interference is destructive, and the intensity of the final signal is minimum (zero in the ideal case). By varying the frequency of the emitted signal, Appleton could locate two successive maxima and, without knowing *a priori* the number of wavelengths involved, deduced the distance of the reflecting layer. He found it at about 90 km. Being aware that reflections can come from buildings or hills around, he developed another technique that allowed him to determine the angle under which the measured signal arrived and confirmed that the reflection occurred at an altitude of 90 km. Continuing with the experiments, Appleton sought to explain how an electromagnetic wave can be reflected by the atmosphere. Using the laws of electromagnetism, he proposed a theory called "magneto-ionic" which accounts for the propagation (including the reflection) of a radio wave by an ionised layer consisting of free electrons and ions where there is a magnetic field in the background (the geomagnetic field). This discovery was not accepted without opposition and the editor of the journal *Nature* felt obliged to warn, at the top of the 1925 article by Appleton and Barnett, that "the editor is not responsible for the opinions expressed by his correspondents". Appleton obtained the Nobel Prize in Physics in 1947 "for his investigations on the physics of the upper atmosphere, in particular for the discovery of the so-called Appleton layer". This nomination superbly ignored Robert Watson-Watt, an engineer who greatly contributed to the development of radar and who introduced the currently used name of that layer – the "ionosphere" in 1926.

The property of the ionised upper atmosphere to reflect radio waves of certain frequencies was used in 1925 to develop a new instrument, the ionosonde. By varying the frequency of the emitted wave in the megahertz range it was realised that the reflection does not occur at the same altitude. Signals at 1 MHz are reflected from about 90 km altitude. But the more the frequency was increased, the higher the reflecting altitude. At 2.8 MHz, the reflection occurred at about 180 km. And then there was a huge jump. Suddenly the waves reached altitudes ranging from 200 km at 3 MHz to 400 km at 9 MHz. However, all previous theoretical developments had shown that this reflection depends only on one parameter: the electron density at the altitude of reflection. This technique gives two pieces of information: the round trip time between the emission and the reception of the signal which provides an effective altitude of the reflecting layer, and the frequency of the emitted wave which allows one to calculate the density of free electrons. By varying the frequency of the wave, it is possible to probe the ionospheric medium because for each altitude there corresponds a value of the electron density. We quickly understand that in the ionosphere, there are two zones with electron density peaks. The first, between 100 and 200 km, with densities of about one hundred thousand free electrons per cubic centimetre, is called the E layer. Above, the density increases to a million electrons per cubic centimetre between 250 and 400 km. This is the F layer. This technique, relatively simple and inexpensive, is still used today.

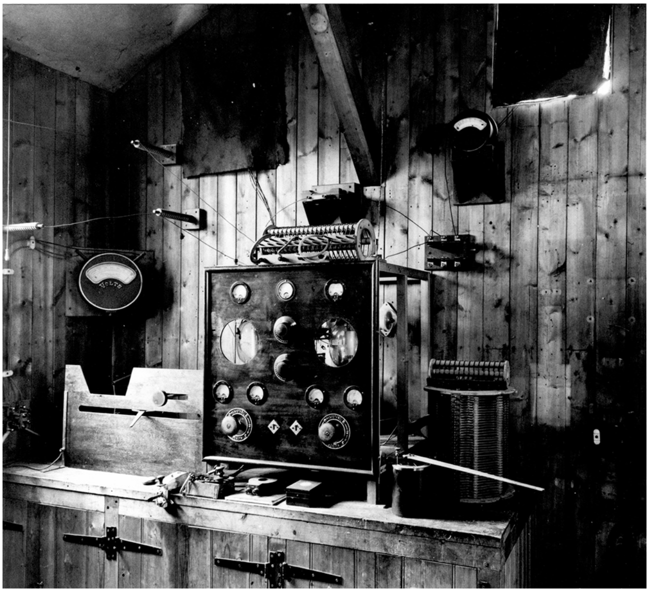


Fig. 2.9 – Ionosonde of the 1930s. This device was connected to a transmitting/receiving antenna.

Where does this ionosphere come from and why does it reflect radio waves at specific altitudes? For this, Sydney Chapman (1888–1970) comes on the scene again. In a lecture given to the Royal Society in 1931, he proposed an interesting theory of ionisation of the upper atmosphere by solar extreme ultraviolet radiation. Although this radiation was purely hypothetical at that time, with a typical wavelength of about ten nanometers, it would have sufficient energy to ionise the upper atmosphere, *i.e.* to tear off at least one electron from the main constituents, atoms and molecules. Today, we know that this solar extreme ultraviolet radiation indeed exists and originates from the one million-degree hot solar corona. Chapman's description implies a typical altitude profile of electron density and reproduces well the maximum observed around 250 km altitude with the first ionosondes.

Nomenclature for the Atmospheric and Ionospheric Layers

At a time when research on natural plasmas was in full swing and the close interactions between the various regions of space and atmospheric layers were being realised, Sydney Chapman (again!) remarked in a 1950 article that the notion of the 'upper atmosphere' was practical but not very precise, and he took stock of the nomenclature of atmospheric layers. He thus extended the work of Teisserenc de Bort (1855–1913, see chapter 1). In particular, he deplored the fact that the term stratosphere was used for any altitude above the troposphere and introduced, in order to extend the classification based on thermal considerations, the names *mesosphere* to designate the zone in which the temperature of neutral species, as was assumed at the time, peaked; and then *thermosphere* for the neutral gas above about 100 km, where the temperature was supposed to rise sharply. He notes in his paper that the term thermosphere should refer to the layer up to the next temperature minimum if it exists. Rockets and satellites later indicated that this hypothetical minimum in the upper part of the thermosphere does not exist.

Similarly, the ionosphere is the part of the upper atmosphere composed of free ions and electrons. It is divided into layers according to the density of free electrons found in it and by the properties of the plasma. The highest layer, which is present also during nighttime, is the F layer with a peak in electron density. It is enhanced during daytime by solar extreme UV radiation. Below that, between about 90 and 120 km we found the E layer, created from ionised molecular oxygen (O_2) by soft X-rays and far UV solar radiation. The D layer, which is present at around

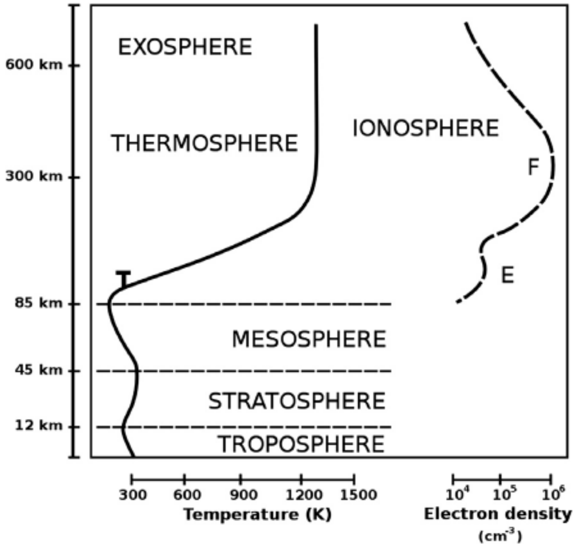


FIG. 2.10 – Different atmospheric layers (as a function of temperature of neutral species) and ionospheric layers (as a function of electron density).

70–90 km altitude, is not depicted in the enclosed figure as its electron density is at least an order of magnitude lower than for the E layer. In the E layer, most of the ionisation comes from Lyman alpha hydrogen UV radiation, but this can be enhanced considerably by cosmic-ray particles or very energetic X-ray photons. Some of the D and E layer ionisation also comes from the decay of meteors.

The First Idea on the Variable Earth's Magnetosphere

In 1930, Chapman and his student Vincenzo Ferraro (1907–1974) announced in a letter published in Nature that they were about to propose a new theory to explain magnetic storms. This explanation, which appeared in a series of articles published over the next three years, was based on the interaction of the gas ejected by the Sun with the Earth's magnetic field. They attempted to explain the variations in the magnetic field on the ground (and the associated equivalent currents) by the distribution of charges at the interface between the flow of particles from the Sun (not yet called the solar wind) and the geomagnetic field. Chapman and Ferraro felt that the interaction between this flow of particles and the geomagnetic field must lead to a deformation of the latter, although in their papers this remains rather vague: they represented the interface alternately as an ellipsoid and a plane.

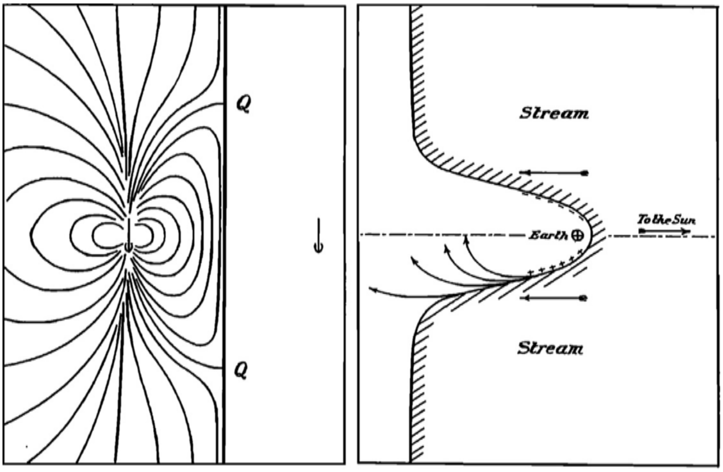


Fig. 4—Meridian section

Fig. 5—Equatorial section

Fig. 2.11 – The magnetosphere after Chapman and Ferraro (1931). On the left is a midnight meridian section showing an interface between the magnetosphere and the external environment; on the right is the shape of the magnetosphere in the equatorial plane with the flow of particles of solar origin across it.

They predicted the existence of an electric current at the interface, due to the separation of negative charges (electrons) which they said tend to go to the morning side and positive charges (ions) which tend to go to the evening side. This creates a current flowing from morning to evening (or from west to east) and induces a

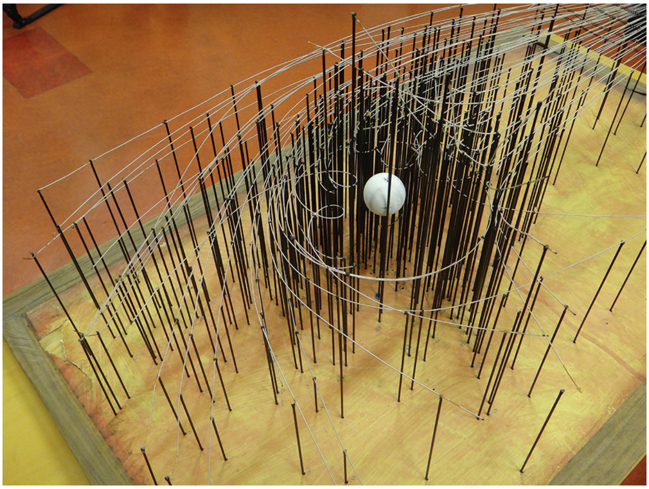


Fig. 2.12 – To model the trajectory of electrons in the Earth's magnetic environment before the digital age, Carl Stormer used umbrella ribs. Using white metal wires, he drew the path of the particles, calculated from Maxwell's equations. The model is at the Tromsø Auroral Observatory in Norway (Credit: Jean Lilensten, CNRS – IPAG).

magnetic variation towards the north observed on the ground during the triggering of magnetic storms (Storm Sudden Commencement). This current does indeed exist; it will later be called the Chapman-Ferraro current. With this, Chapman and Ferraro laid the foundations for solar wind-magnetosphere coupling and contributed to opening the way to space weather and space climate physics, as will be shown in the next chapter.

References

Alfvén H. (1988) Memoirs of a dissident scientist, *Am. Sci.* **76**(3), 249–25.

Appleton E.V. (1925) The propagation of radio waves over the earth, *Nat.* **115**(2889), 382. https://doi.org/10.1038/115382d0.

Appleton E., Barnett M. (1925) Local reflection of wireless waves from the upper atmosphere, *Nat.* **115**(1925), 333–334. https://doi-org.gaelnomade-1.grenet.fr/.

Barnett M.A.F. (1974) The early days of ionosphere research, *J. Atmos. Terr. Phys.* **36**, 2071–2078. https://doi.org/10.1016/0021-9169(74)90138-X.

Chapman S. (1931) Some phenomena of the upper atmosphere, *Bakerian lecture* given June 25, 1931, *Proc. R. Soc. A.* **132**, 353. https://royalsocietypublishing.org/doi/pdf/.

Durham I.T. (2006) Rethinking the history of solar wind studies: Eddingtons's analysis of comet Morehouse, *Notes Rec. Royal. Soc.* **60**, 261–270.

Egeland A., Burke W. J. (2006) Kristian Birkeland: The First Space Scientist, Springer.

Hughes S. (2012) Catchers of the light: The forgotten lives of the men and women who first photographed the heavens. ArtDeCiel Publishing, pp. 202–223. ISBN 978-1-62050-961-6.

Lakhina G. S., B.T. Tsurutani (2018) Chap. 7 – Supergeomagnetic Storms: Past, Present, and Future, in Extreme Events in Geospace (ed. N. Buzulukova), Elsevier.

Lindemann F.A. (1919) LXX. Note on the theory of magnetic storms, *Philos. Mag. Ser.* **6**, 669–684.

Lyot B. (1930) La couronne solaire étudiée en dehors des éclipses, *Bull. Astronomique* **6**, 305–316.

Lyot B. (1935) Quelques observations de la couronne solaire et des protubérances en 1935, *L'Astronomie* **51**, 203.

Marchand E. (1904) La perturbation magnétique du 31 octobre 1903, Bulletin de la Société Astronomique de France et Revue Mensuelle d'Astronomie, de Météorologie et de Physique du Globe **18**, 34–36.

Mazotto D. (1905) Télégraphie sans fil, Dunod.

Obridko V.N., O.L. Vaisberg (2017) On the history of solar wind discovery, *Sol. Syst. Res.* **51**(2), 165–169.

Stern D. P. (1989) A brief history of magnetospheric physics before the spaceflight era, *Rev. Geophys.* **27**(1), 103–114, https://doi.org/10.1029/RG027i001p00103.

Stern D. P. (1996) A brief history of magnetospheric physics during the space age, *Rev. Geophys.* **34**(1), 1–31. https://doi.org/10.1029/95RG03508.

Chapter 3

The Time of Complexity: The Earth

In the middle of the twentieth century, many obscure points in the understanding of the Sun-Earth connections remained and led to interesting scientific controversies. Let us be clear about what a scientific controversy is. It is an open question debated by the scientific community, leading to at least two unclear and *a priori* incompatible explanations. It has nothing to do with media controversies, such as those that can be found today on subjects on which there is a consensus among scientists… An event which catalysed research, removed controversies and, what is so important for the exploration of the Earth's environment and beyond, launched the space race was the International Geophysical Year.

The Beginnings of the Space Age in Europe

Since space weather and space climate are so dependent on space exploration, it is worth recounting how Europe entered the space age.

First of all, the International Geophysical Year in 1957 and 1958 gave a boost to all scientific activities concerning the upper atmosphere and the means of sounding it, including rockets. Also, the Soviets struck a blow by sending the first artificial satellite, Sputnik, into orbit on 4 October 1957, which shook everyone; the space race was launched. We no longer spoke of space exploration but of the conquest of space. This change in terminology is very revealing of the rivalry between the Americans and the Soviets; the Cold War will also be a 'scientific war' that will boost innovation.[1] Charles de Gaule, who was once again leading France, believed in the potential of rockets. He wrote to his Soviet counterpart Khrushchev in 1958:

> "We are in the century of rockets and aircraft, and humanity should not be deprived of them."[2]

---

[1] Rasmussen Anne, Guerres et sciences, in Histoire des sciences et des savoirs, sous la dir. D. Pestre, 2015.

[2] Note dated 28 june 1960, Charles de Gaulle*, Lettres, notes et carnets (1958–1960)*, Paris, Plon, 1985, p. 371.

DOI: 10.1051/978-2-7598-3610-9.c003

FIG. 3.1 – Cover of the magazine Sciences et Avenir, April 1952.

Setting up the national space agencies was accelerated. But we can imagine that the creation of an intergovernmental agency like the European Space Agency (ESA) did not happen overnight. In fact, it took more than 15 years to create the current European Space Agency. The first discussions on this subject date back to the late 1950s. As with French space activities, development at the European level benefited from the military heritage of rockets and the International Geophysical Year. European scientific initiatives such as Euratom and CERN also set an example and paved the way in pooling efforts, which was the only way to compete with the superpowers of the United States and its fledgling NASA, which came into being in 1958, and the USSR.

The history of the creation of ESA, which is extensively detailed in a two-volume book,[3] went through some key stages and other embryos of the agency. Cosmic ray

---

[3] Krige, J., and A. Russo, A history of the European Space Agency 1958–1987, SP 1235, 2000.

physicist Pierre Auger (1899–1993) who was instrumental in setting up the French space agency with his Italian physicist colleague and friend Edoardo Amaldi (1908–1989) drawing inspiration from the CERN model organised an intergovernmental meeting in Paris on 23 and 24 June 1960, bringing together representatives from ten European countries: Belgium, Denmark, France, Germany, Italy, the Netherlands, Norway, Sweden, Switzerland and the United Kingdom. This meeting was to be known as the European Space Research Study Group (ESRG). The conclusion of this meeting was the creation of a European Preparatory Commission for Space Research (COPERS). COPERS, which met for the first time in Paris in March 1961, aims to discuss and negotiate the modalities for the creation and operation of a European space research agency. ESRO, the European Space Research Organisation, was created on 14 June 1962. The ten Member States committed themselves for a period of eight years to a joint programme which was limited to the development of launchers and satellites. In parallel with ESRO, a body dedicated to launchers, ELDO (European Launcher Development Organisation) was created in April 1962.

It became clear that ESRO and ELDO did not have the means to fulfil their ambitions, and that their activities were threatened. As European space activities cannot be allowed to fail, representatives of the governments involved and the heads of the European space centres of the day met in Brussels on 15 April 1975 and signed the founding texts of the current European Space Agency.

Today, ESA has 22 Member States: Austria, Belgium, Czech Republic, Denmark, Finland, France, Germany, Greece, Hungary, Ireland, Italy, Luxembourg, the Netherlands, Norway, Poland, Portugal, Romania, Spain, Switzerland, Sweden and the United Kingdom. Some other European countries have agreements and privileged relations with the agency. Finally, Canada (which NASA has never accepted into its ranks) participates in some programmes as an associate, as do the remaining European countries. The agency employs about 2500 people at eight sites: in Germany, in the United Kingdom, Belgium, Spain, France, the Netherlands and Italy.

The International Geophysical Year

In 1950, renowned scientists such as Sydney Chapman and James Van Allen (1914–2006) proposed to organise an International Geophysical Year, based on the same model as the International Polar Years of 1882–1883 and 1932–1933. The aim will be to study all related fields of geophysics, in a spirit of international cooperation and mutual aid. The International Geophysical Year was scheduled from July 1957 to December 1958. For the occasion, observation stations were set up in key regions such as the poles. This was the case for the French Antarctic base of Dumont d'Urville, but also for the Soviet Vostok, British Halley and American Amundsen-Scott bases, which are all still operational today.

But to better understand our space environment and the phenomena that occur there and the effects of which we feel on the ground, we need to go into space and measure the properties of the environment in situ. In France, the Véronique sounding rocket programme, which was already well advanced, was being improved

FIG. 3.2 – Véronique rocket at the test bench in Vernon (Normandy) around 1960 (Credit: LRBA).

to reach an altitude of 230 km in its International Geophysical Year version to explore the upper atmosphere.

In the United States, the Vanguard rocket project was launched in 1955 to send exploration satellites into space. It should be borne in mind that at that time, in the middle of the Cold War, the United States and the USSR were battling it out in all areas. The space race was no exception. The Soviets were making great strides; faster than the rest of the world thought. On 4 October 1957, the Sputnik 1 satellite was launched by an R-7 rocket from the now-famous Baikonur base in Kazakhstan. Sputnik 1 was a small sphere 58 cm in diameter weighing about 83 kg. It was placed in an elliptical orbit with an apogee of 950 km and a perigee of 230 km. Its orbital period is 98 min. Sputnik 1 was only a technological test, it had no particular function except for the emission of a radio signal at regular intervals, proof of the success of the launch and orbital insertion. Sputnik 1's life span was short, since on 4 January 1958 it disintegrated in the atmosphere: at such a low perigee, a satellite is slowed down by the upper atmosphere, however tenuous it may be, and requires its orbit to be raised by a propulsion system; otherwise, it will inexorably fall into the denser atmosphere and burn up.

For the Soviets, Sputnik 1 was just a test; in the United States, where the news was very unpopular, it was an earthquake: not only had the Soviets stolen the technological limelight and gained a head start in the space race, but their ability to send rockets into space implied that they were also capable of sending intercontinental missiles to anywhere on the Earth's surface, including the United States!

The Hunt for the Radiation Belt

A major scientific question and therefore an active research topic in the 1950s was the possibility of a radiation belt around the Earth. The theories of that time – which proved to be correct – predicted that highly energetic, electrically charged particles (electrons, protons) would be trapped by the Earth's internal magnetic field. Moreover, the first laboratory experiments on the magnetic confinement of plasmas tend to confirm this theory: in a dipolar magnetic field – with a simple magnet having a north pole and a south pole – particles do have the property of bouncing between two regions of intense field called magnetic mirrors. Does this phenomenon occur in the Earth's environment? To find out, scientists need to send a particle sensor into space.

The Soviets repeated a successful launch on 3 November 1957 with Sputnik 2 which, unlike its predecessor, carried a radiation detector. High radiation levels were measured but went unnoticed. Above all, the Soviets refused to communicate their data to foreign scientists who could have interpreted the results. These results would likely have led to the discovery of the radiation belt.

The US response to Sputnik 2 was hurried. The Vanguard rocket that was to launch the Explorer 1 satellite burned up on the launch pad. The team of the famous German engineer Wernher von Braun (1912–1977), the father of the German V2s, hastily adapted a Jupiter C-type ballistic missile and the Explorer 1 satellite, also built in haste, was launched on 31 January 1958. It carried a Geiger counter supplied by Van Allen's team but no data recorder, so the only data recovered was those received by radio at ground stations as the satellite passed over them. But the onboard instrument sent some curious data: at locations where high levels of radiation were suspected, the counter read zero. In fact, it was later realised that the instrument was saturating due to the large flux of energetic particles. Explorer 2 was a failure. It was not until Explorer 3, launched on 26 March 1958, that the famous radiation was measured – and recorded – correctly, and the presence of a radiation belt was confirmed. Ernie Ray, a researcher on Van Allen's team, was astonished and said: "My God, space is radioactive!" This region was named the Van Allen belt.

It should be understood that Geiger counters measure a level of radiation but do not identify the source of this radiation, *i.e.* they give no indication of the flux or energy of the particles that cause it. These particles remained a mystery. It was assumed at the time that they were electrons, but the flux required to account for

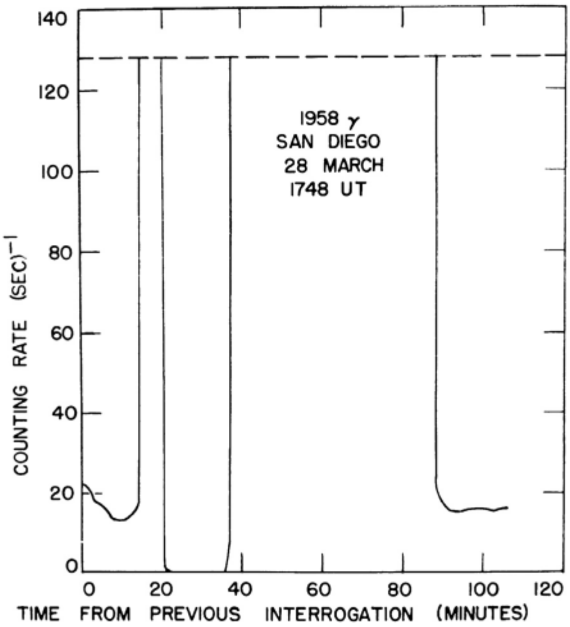


Fig. 3.3 – The first record of radiation belt detection by the Geiger counter onboard Explorer 3. The figure shows the number of 'hits' (detections) as a function of time. Between 20 and 40 min, the detector reaches the saturation point of 128 counts per second and indicates zero (Figure taken from Van Allen's 1959 paper).

the measurements, of the order of 10^8 particles per cm^2 per second, was considered too high. It turned out that this inner belt was mainly populated by protons, which were more energetic.

In December 1958, the Pioneer 3 probe on its way to the Moon (which it did not reach) reached an altitude of about 102,000 km and its Geiger counter indicated that it was passing through a second radiation belt, known as the outer belt, at a distance of between 6 and 8 Earth radii.

At the time, the so-called 'leaky bucket' theory explained that the auroras observed along the auroral oval were produced by particles from the outer radiation belt escaping from it (by crossing the magnetic mirrors). Van Allen writes in a 1959 paper:

> "We propose that the radiation belt constitutes a reservoir from which the escape of particles is the direct cause of the visible aurora. Furthermore, we suggest that the solar wind fills the reservoir from time to time, working its way through the layers of the Earth's magnetic field when its density is sufficiently great, and then becomes trapped in the field."

It was later realised that the particle energy of the radiation belts was too great to explain the aurora. However, the idea of associating the visible polar auroras directly with the Van Allen belts is still persistent today.

The Magnetosphere and Solar Wind are Revealed

The United States continued the Explorer programme, while the Soviets also contributed to the exploration of the space environment with the Luna probes. Without giving an exhaustive account of the programmes, we would like to highlight a few remarkable results that have contributed to a better understanding of the Earth's space environment.

To designate the Earth's magnetic environment, the term *magnetosphere* was proposed by Thomas Gold (1920–2004) in 1959.

Also in 1959, the Luna 1 probe, which carried an ion trap, and a simple particle detector, measured the solar wind and confirmed its existence. The presence of this flow of particles from the Sun was confirmed a little later by measurements from the Luna 2 and Luna 3 probes and by Venera 1 in February 1961, which was sent to Venus but soon stopped transmitting.

Luna 2 also revealed a layer of low-energy plasma in a toric zone below 65° latitude extending from the upper ionosphere to about 4–5 Earth radii in the plane of the equator. It was called the plasmasphere by Don Carpenter (1928–2019) in 1966, but its presence had been suspected earlier by scientists working on radio wave propagation. Indeed, they sometimes heard parasitic sounds of varying frequency, which they called whistlers. These whistlers are very low-frequency electromagnetic waves which propagate in the plasmasphere but are generated by tropospheric thunderstorms.

On 25 March 1961, Explorer 10 was launched and placed in a highly elliptical orbit with an apogee of 28 Earth radii in the direction of the night side. The instruments on board indicated that the satellite had stealthily passed into the solar wind and made its first velocity measurement.

Explorer 12 was sent on 16 August 1961 to probe the dayside (facing the Sun) of the Earth's magnetic field. The probe passed through a region that seemed to separate the Earth's magnetic environment from the solar wind (in fact, it was later discovered that the panorama was once again a little more complex and that this layer was still made up of the slow-moving solar wind; it will be called *magnetosheath*). The boundary layer that delimits the region in which the Earth's magnetic field dominates is called the *magnetopause.*

However, a doubt remains. The measurements were only made over small time intervals. Is the solar wind always present or is it only intermittently ejected by our star? Launched on 27 August 1962, the Mariner 2 probe en route to Venus settled the question by sending back months of data showing that the solar wind continuously bathes interplanetary space and that it flows at variable but always supersonic speeds. These measurements put an end to a more than 60-year-old controversy concerning the existence of the solar wind and the interplanetary magnetic field and proved Størmer, Parker and others right.

The Explorer 18 data (renamed Interplanetary Monitor Platform, IMP 1) subsequently show a periodicity of 27 days (the Sun's average rotation period) in the solar wind properties, confirming the hypotheses of Alfvén and Parker who had understood, even before the in situ measurements, that a radial solar wind emitted by a rotating Sun should result in an interplanetary magnetic field in the shape of an Archimedean spiral (Parker spiral). Explorer 18 also observed a shock zone before the magnetopause: this is the bow shock through which the supersonic solar wind is slowed to subsonic plasma speeds. The probe also explored the tail of the magnetosphere.

With the proliferation of scientific satellite missions in which the United States asserted its prominence, the different regions of the magnetosphere were revealed.

In 1971, the polar cusps, very particular regions of the magnetosphere were observed – not surprisingly, since they were implicitly predicted by Chapman and Ferraro's model as the regions of direct entry of the decelerated solar wind (of the magnetosheath plasma, in fact) into the magnetosphere and the ionosphere. From the magnetic point of view, these are the regions of magnetic field lines (called open) connected to the interplanetary magnetic field that separate the closed field lines on the dayside and the lines that form the envelope of the magnetospheric tail.

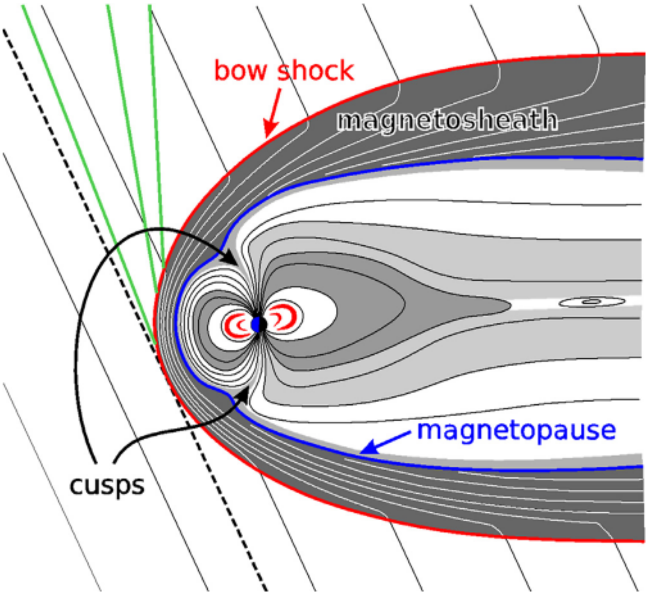


FIG. 3.4 – Cross-section of the magnetosphere in the midnight-meridian plane with some of its regions. The black lines show the interplanetary magnetic field (IMF) direction. If magnetic field lines point towards the bow shock they heavily oscillate near it (the case in the upper part). Waves are due to reflected particles from the bow shock (in green are from left to right depicted boundaries of first upstream occurrences of electron, ion field-aligned beam and ion diffusion regions) as particles travel easily along the mean IMF direction. At the bottom part, reflected particles predominantly just after reflection VxB drift back to the bow shock. In fact, this representation hides all the complexity of the 3D, the dynamics and the fact that regions are partly co-located.

The polar cusps were observed at low altitudes by the Canadian satellite ISIS-1 and at higher altitudes by the NASA satellites IMP-5 and OGO 5.

Thus, in just fifteen years, probes have explored and discovered the Earth's space environment, and begun to clarify centuries of contradictory and mysterious observations. The magnetosphere has proved to be a dizzyingly complex and therefore rich source of information for physicists. It is a natural laboratory for the study of plasmas and a whole 'zoology' of waves that propagate in it. With the data accumulated and the knowledge acquired, scientists are finally in a position to provide answers to questions that are sometimes centuries old and to confirm or refute the hypotheses put forward.

Is the Magnetosphere Closed or Open?

An important point in understanding the coupling between the solar wind and the magnetosphere is the entry of particles of solar origin into the magnetosphere. On the one hand, Hannes Alfvén's theory of magnetohydrodynamics predicts that two plasmas linked to two distinct magnetic fields (solar and terrestrial) cannot interpenetrate. However, observations suggest the opposite: auroras are more intense during peaks in the Sun's activity and satellite measurements show similar properties of particles on either side of the magnetopause. So, is the magnetosphere the protective cocoon it has been described as? Is the magnetopause such a hermetic boundary layer to the solar wind? Answering these two questions is not easy. To the first order, it can indeed be said that the terrestrial (or planetary) magnetosphere protects from the solar (or stellar) wind. The planetary magnetic field acts as a shield to the flow of the solar wind around the magnetosphere, leaving the interior of the magnetosphere almost 'empty'. At least, with a much lower density of particles, between 100 and 1,000 times lower. However, it is also true to say that there is leakage. Particles from the solar wind do manage to cross the magnetic barrier of the magnetopause. But how? This has been the subject of endless discussions and debates in papers and articles since the 1960s.

At that time, two models were in conflict: the closed magnetosphere model, which allows only 'viscous' interactions at the magnetopause between the solar wind plasma and the magnetospheric plasma, and the open magnetosphere model, proposed by James W. Dungey in 1961, which predicts an interconnection between the interplanetary magnetic field (which comes from the Sun) and the geomagnetic field. The latter model has a major advantage: not only does it allow for the entry of solar wind particles into the magnetosphere, but it also accounts for a series of crucial observations: geomagnetic activity in general is greater when the interplanetary magnetic field is oriented towards the south, *i.e.* in the opposite direction to that of the geomagnetic field. The idea then began to emerge that when two magnetic fields are opposed, they could locally cancel each other out and reconfigure themselves. This process has been demonstrated in natural and laboratory plasmas and has been called *magnetic reconnection* – we will come back to this later. This process, although attractive, is not self-evident. Indeed, it locally violates the laws of ideal magnetohydrodynamics, including the freezing-in of the magnetic field in the plasma.

An alternative explanation was proposed in the 1970s by the Belgian Joseph Lemaire: impulsive penetration, which states that during solar wind pressure pulses (due to a sudden increase in particle density and/or speed) the solar wind can pass through the magnetic shield and into the magnetosphere. This explanation was favoured for a while by those who were opposed to magnetic reconnection and, even though observations seem to give it credence under certain conditions, it did not really succeed in gaining acceptance.

The Dynamics of the Magnetosphere

Successive magnetospheric missions have established that the magnetosphere is a highly dynamic medium, constantly changing according to external conditions (solar wind speed and density, amplitude and orientation of the interplanetary magnetic field). As it is useful to discuss some manifestations of these dynamics in order to understand all the issues and needs of space weather and space climate; let us mention a few.

The first is the size of the magnetosphere, which is constantly adapting to the pressure of the solar wind. If the solar wind is tenuous or slow, the magnetosphere is voluminous; conversely, when the solar wind is dense or fast, it compresses the magnetosphere and its dimensions are reduced. Thus, the subsolar point, which is the nose of the magnetosphere in a way, can vary its distance from the Earth between about 5 and 15 Earth radii. We will see in chapter five the consequences of these variations. A direct corollary: since the auroral oval corresponds to the projection into the ionosphere along the magnetic field lines of the zones of the magnetosphere most populated by charged particles, if the magnetosphere contracts, the auroral oval will widen and the auroras will be visible from the lower latitudes. Two important parameters for predicting the state of the magnetosphere are the velocity and density of the solar wind.

The second is the role of the orientation of the interplanetary magnetic field, on which the coupling state between the solar wind and the magnetosphere depends. We have seen that according to Dungey's model, when the interplanetary magnetic field is oriented southwards, it is antiparallel to the geomagnetic field and the process of magnetic reconnection can occur: locally, the two fields cancel and reconfigure themselves, creating 'open' field lines on the solar wind. Matter and energy from the solar wind can then enter the magnetosphere. Moreover, the stronger the interplanetary magnetic field, the more effective the coupling. On the other hand, when the interplanetary magnetic field is oriented towards the north, the coupling will not be as efficient and the energy and particle transfer will occur significantly less. Here it is the strength and orientation of the interplanetary magnetic field that is important.

The third manifestation of magnetospheric dynamics is the circulation of plasma in the magnetosphere and ionosphere, which is incorrectly called convection. Indeed, the fact that the magnetosphere is open results in plasma flowing from the dayside

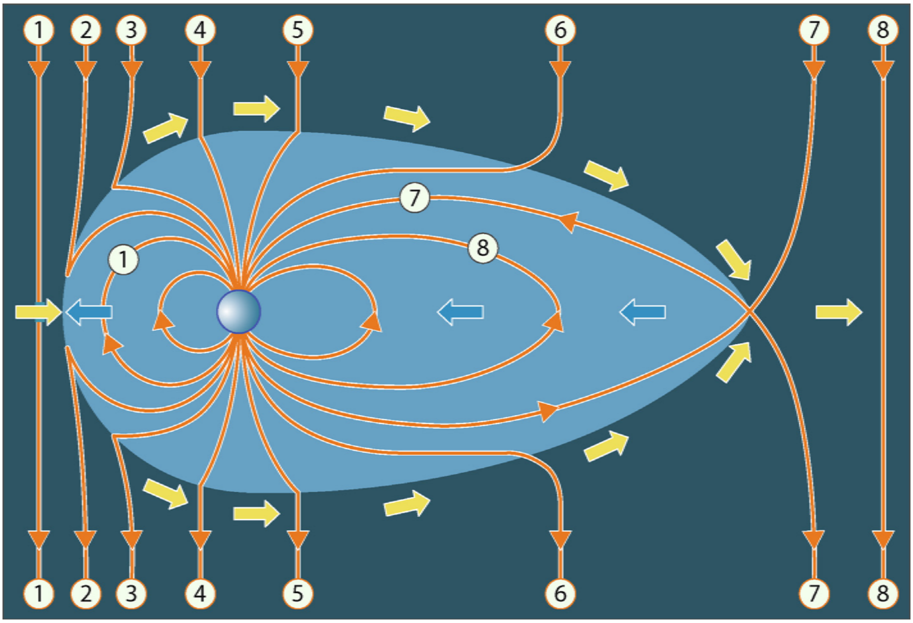


FIG. 3.5 – Visualisation of the consequences of magnetic reconnection and magnetospheric convection in several steps and in a midnight-meridian section. Step 1: a southward directed interplanetary magnetic field (IMF) line approaches and reconnects with a geomagnetic field line. Steps 2–6: the two open magnetic field line halves are driven by the solar wind towards the tail of the magnetosphere (meanwhile, the plasma in the ionosphere at the foot of these field lines is transported through the polar zones from the dayside to the nightside). Step 7: The two open line halves are connected to give a closed line and an open line. Step 8: The closed line moves towards the Earth under the combined action of the magnetic and electric field while the open line, disconnected from the Earth, moves away with the solar wind. (Figure after Baumjohann and Treuman, 1996).

to the nightside at high latitudes and from the nightside to the dayside at lower latitudes. This plasma flow is also very sensitive to external conditions. It is important because it drives the movement of plasma in the ionosphere, which interacts with the neutral atmosphere, which would be otherwise not affected by this convection. The interactions – we can say friction for simplicity's sake – between ionised and neutral particles in the upper atmosphere lead to a heating of the medium.

Also, the motion of charged particles causes electric currents to flow in the magnetosphere. We will mention here only three of them. Two flow perpendicular to the magnetic field lines: the Chapman-Ferraro current which flows from west to east along the magnetopause on the Sun's side and the ring current which flows from east to west in the inner magnetosphere, in a band between 3 and 8 Earth radii; and the currents that flow along the magnetic field lines (the so-called aligned or Birkeland currents) which close the Chapman-Ferraro and ring currents in the high latitude ionosphere. The intensity of these currents depends highly on the state of coupling between the solar wind and the magnetosphere. The same applies to the ionosphere, where current systems develop inducing geomagnetic variations and induced currents at ground level (see chapter 5).

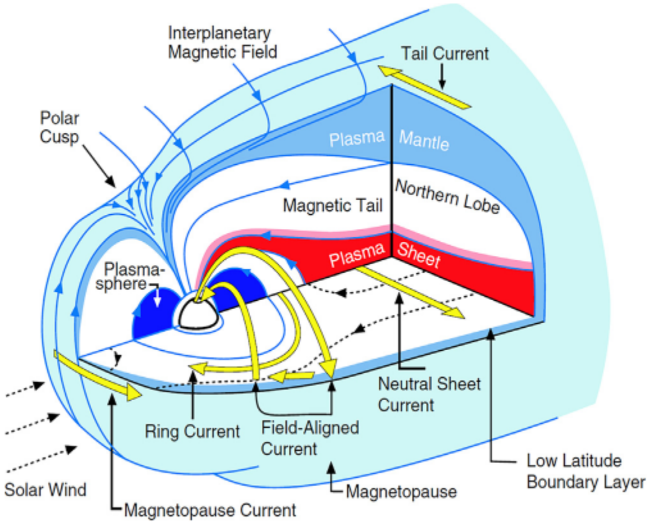


Fig. 3.6 – Three-dimensional model of the magnetosphere. The magnetopause and magnetopause boundary layer appear in light blue and darker blue, respectively. In yellow are shown different electric current systems – the magnetopause "Chapman-Ferraro" and ring currents; the field-aligned "Birkeland" currents; the magnetospheric tail and neutral sheet currents. Credit to "Magnetopause and Boundary Layer", De Keyser *et al.*, 2005.

It can be seen that the simple two-dimensional cross-section of the magnetosphere as presented a few pages back is more than simplistic!

Magnetic Reconnection

The phenomenon of magnetic reconnection, which we have already mentioned, is worth looking into because it is a fundamental process in natural or laboratory plasmas. It allows the magnetic field to give up its energy to matter, and thus to heat it and/or accelerate it. Let us see where it comes into play in the context of the Sun-Earth relationship in four examples.

Solar chromospheric flares: when a magnetic loop emerges from the solar surface, it may lengthen and tighten at the bottom, so that magnetic field lines of opposite directions are brought into contact. The phenomenon of magnetic reconnection then kicks in to disconnect a plasma bubble that will be ejected. This is often accompanied by powerful UV and X-ray emissions.

The heating of the solar corona: the processes that heat the solar corona to some 2 million degrees while the Sun's surface is only about 5,500 °C still remains a

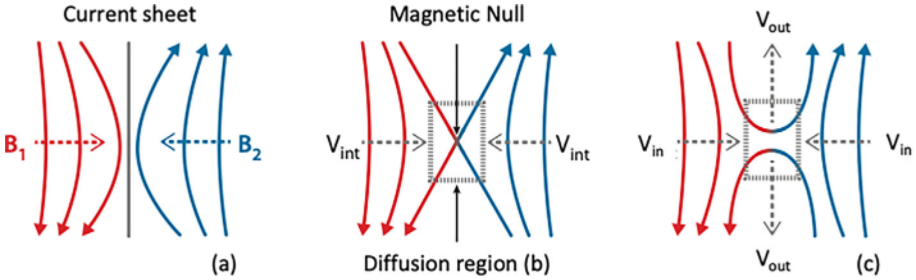


FIG. 3.7 – Three stages of 2-D depiction of magnetic reconnection: (a) two regions of opposing magnetic fields approach each other, separated by a layer of current; (b) locally, the magnetic field is cancelled and in the so-called diffusion region, the field lines reorganise; (c) the reconnected magnetic field lines (half red and half blue) are accelerating plasma carried within them on either side of the reconnection site.

mystery and a major scientific challenge. Among the processes invoked, magnetic reconnection in flares (as described above) or on a much smaller scale could explain, in part, this heating.

Plasma jets in the solar wind: MHD explains that the plasma in the solar wind and the interplanetary magnetic field are frozen into each other and therefore propagate at the same speed. However, the ejection speed of the solar wind varies over time and a fast wind can catch up with a slower one. At the interface of the two, an overdensity of plasma forms and magnetic fields of opposite directions can be brought into contact. Magnetic reconnection can then occur and generate characteristic plasma jets.

The penetration of the solar wind into the magnetosphere: as we have already mentioned, magnetic reconnection is the number one candidate to explain the penetration of the solar wind (of the magnetosheath plasma in any case) into the magnetosphere. Closed geomagnetic field lines connect to interplanetary magnetic field lines, allowing solar-derived plasma to be on magnetic field lines that are connected at only one end to the Earth.

Geomagnetic Storms and Magnetospheric Substorms

The large magnetic disturbances measured on the ground have been intriguing since the mid-19th century. The link with solar activity has been established and electrical currents are suspected of inducing these disturbances, but where do they flow and when? In fact, two distinct – but often linked – events have been mixed up and confused: geomagnetic storms, which affect the entire magnetosphere-ionosphere system, and magnetospheric substorms (once called polar storms), which occur in the tail of the magnetosphere and are essentially felt only at high latitudes on the night side.

Magnetic storms are the most extreme manifestation of the coupling between the solar wind and its interplanetary magnetic field, and the Earth's space environment. They occur when magnetic reconnection at the dayside magnetopause (mainly driven by a southward interplanetary magnetic field) is efficient for several hours. Magnetic reconnection at the dayside is a key process regulating the energy input from the solar wind into the magnetosphere. During magnetic storms, this energy input leads to the intensification of the ring current and several other current systems (such as field-aligned currents) in the magnetosphere.

Fast and dense solar wind can also compress the magnetosphere which intensifies the electric currents flowing along the magnetopause (Chapman-Ferraro current) and in the magnetosphere (ring current and Birkeland current). Magnetic storms can last several days and affect the Earth's space environment (movement and heating of the thermosphere and ionosphere, disturbances of the radiation belts, etc.) also human technologies, as we will see in chapter 5. To know the magnitude of a magnetospheric storm, one needs to know the properties of the solar wind (density and speed) as well as the amplitude and orientation of the interplanetary magnetic field. Predicting these properties is precisely one of the major challenges of space weather and space climate as described in the following chapters.

Magnetospheric substorms originate in the tail of the magnetosphere. They constitute – still today – one of the most persistent enigmas of magnetospheric physics. Indeed, two models clash: the current disruption model and the distant reconnection line model. The former predicts that the electric current flowing through the tail of the magnetosphere is short-circuited and closes in the ionosphere *via* aligned currents, causing reconnection in the far tail so that the magnetosphere adapts to its new configuration; the latter implies that magnetic reconnection takes place far into the magnetospheric tail first and that this reconnection causes the current to be disrupted and particles to be ejected towards the Earth. Both models lead to more or less the same result: large magnetic disturbances measured at ground level and intense auroras on the night side. In the end, what separates the two models is the chronology of events, and for physicists, this makes all the difference in solving such controversies!

The South Atlantic Anomaly

Magnetic readings from ground stations in the South American regions suggested the existence of an area of remarkably low magnetic intensity; this was confirmed by satellite measurements. The existence of this region called the South Atlantic Anomaly is due to the fact that the Earth's magnetic axis does not pass through the centre of our planet but intersects the plane of the equator about 500 km from the geographic axis towards Asia. The South Atlantic Anomaly is therefore the region on Earth farthest from the magnetic axis and therefore where the magnetic field is weakest. As a corollary to this greater distance from the magnetic axis, the inner Van Allen belt is closer to the ground than anywhere else; for low-altitude satellites, the space environment is more aggressive with higher energy particle flows there.

Dynamics of the Upper Atmosphere

What if we went back a little closer to the Earth? Since the magnetosphere is connected to the neutral upper atmosphere and the ionosphere by the magnetic field

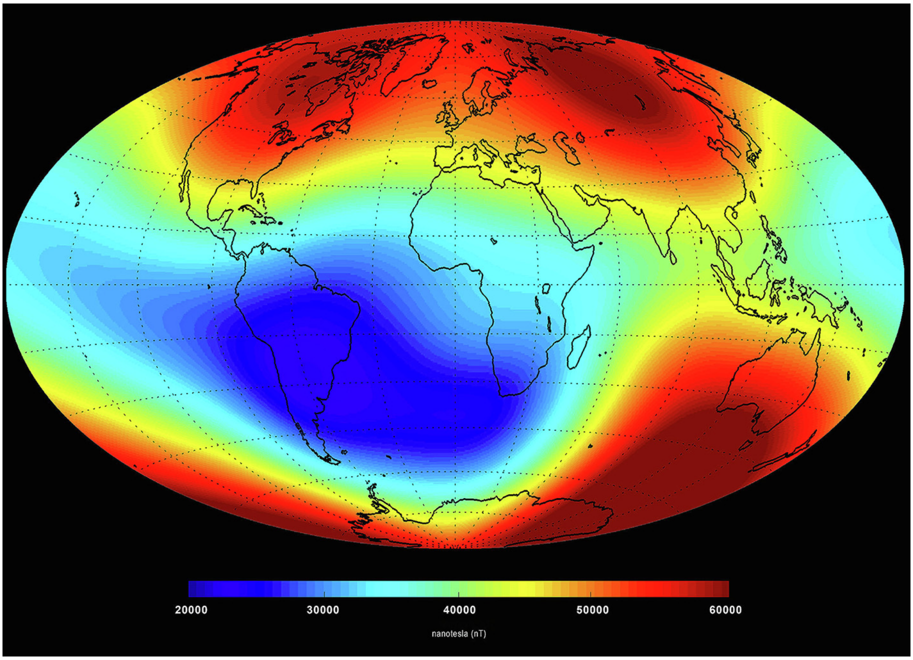


FIG. 3.8 – Ground magnetic field intensity showing the South Atlantic magnetic anomaly in dark blue (Credit: ESA/DTU Space).

lines along which particles, currents and waves circulate, any variability in the former necessarily affects the latter two.

Variability of the Aurora

We saw in the first chapter that the aurora occurs in regions that form two ovals centred on the geomagnetic poles. Why this oval shape? The magnetosphere is more or less full of charged particles depending on the region, and to produce visible auroras (red, green, blue, purple for the most common), particles of particular energies are needed, of the order of one kiloelectronvolt (energy acquired by an electron accelerated by a voltage of 1,000 V): if the energy is too low, the particles do not excite the upper atmosphere; if the energy is too high, the particles go too fast and cross the atmosphere, interacting only moderately with it. It turns out that in the magnetosphere, a reservoir of particles of about 1 keV surrounds the Earth: the plasma sheet. If we extend this plasma sheet along the magnetic field lines in both hemispheres, we get the auroral ovals. The auroral oval is actually therefore centered about 3–5° nightward of the magnetic pole so that auroras occur closer toward the equator during the local (magnetic) midnight.

We have explained that these two ovals contract or expand according to the orientation and amplitude of the interplanetary magnetic field, but how can we

FIG. 3.9 – Polar aurora at Skibotn, Norway, February 2023 (credit: Gaël Cessateur/IASB).

explain the variability of an aurora? When we have the chance to admire them, we realise – apart from the particular case of the so-called diffuse auroras, which do not have a particular shape – that the auroras offer a variety of shapes and movements. Why do they do this? We can give several answers. Firstly, a set of charged particles – a plasma – can be seen as a fluid and it is well known that a fluid moves in a turbulent manner most of the time, so it is not surprising that auroras reflect this turbulence. Also, fluids of very different speeds that are brought into contact see their interface deformed, producing waves; this is called instability. These instabilities can also cause shapes such as ripples and spirals. Finally, electric fields are ubiquitous in the magnetosphere and ionosphere and these highly variable fields deflect the charged particles that create the aurora in equally variable ways.

Variability of Airglow

Other light emissions are less visible – at least to the naked eye – but they occur everywhere on the surface of the globe. This is the airglow of the night sky. Unlike the aurora, they are not caused by the impact of charged particles from the magnetosphere, but by collisions between atoms and molecules, or with ions and electrons still present in the upper atmosphere. When the temperature of the species is high enough, these atoms and molecules collide and transfer energy to each other (this is called collisional excitation); the excess energy is released as light. These night sky glows typically occur at the same altitudes as the aurora. For example, for emissions from oxygen atoms, this is around 100–150 km altitude for green and 250 km for red.

The hotter and denser the thermospheric environment, the more intense the glow. During geomagnetic storms or less intense disturbances, the thermosphere tends to heat up, especially in the auroral zones where the energy input is high, and this creates compression waves that propagate in the thermosphere. These waves appear as striations in the glow of the night sky; from this, we can deduce the

FIG. 3.10 – Astrophotographer Maxime Tessier captures green nightglow on a mountain in Lozère (Credit M. Tessier, https://maximetessierphotographie.fr).

direction and speed of propagation of these waves, which generally go from high latitudes towards the equator. On the same principle, one can also observe isolated disturbances propagating in the thermosphere (Travelling Atmospheric Disturbances) and their ionised counterpart (Travelling Ionospheric Disturbances). Dense plasma bubbles moving through the ionosphere can also be observed by observing the night sky glow.

Cosmic Rays

The solar wind reaches the Earth at an average speed of 370 km per second. However, large solar events launch solar energetic particles (SEPs) which travel much faster and historically they were called solar cosmic rays. Such a change is not without impact on the Earth's space environment, and it has taken many decades to measure it. This very fast, quasi-relativistic wind has been given the unfortunate name of cosmic rays: it is not radiation, but particles. The discovery of these particles, which were already predicted at the beginning of the 20th century, earned Victor Franz Hess (1883–1964) the Nobel Prize in Physics in 1936.

The Sun is a very common type of star in our galaxy. All such stars produce a stellar wind (similar to solar wind) and, sporadically, more energetic particles, called solar energetic particles in the case of the Sun. The stellar wind is stopped sooner or later by collision in interstellar space. The number of stars is so great that our solar system is in fact permanently bathed in a shower of energetic particles arriving from all directions, but most abundantly from the centre of the Milky Way. The high energy cosmic rays with speeds close to that of light are arriving from more exotic sources able to accelerate them so significantly – like black holes, and hypernovae, most of them supposedly from the active cores of the galaxies, so-called active galactic nuclei.

Some of these fast, but negligibly interacting particles like neutrinos pass through the Earth mostly without a hitch. Since you have been reading this chapter, it is likely that some of these particles have passed through without you noticing. But energetic charged particles from out of the solar system can reach the Earth only if they have about five orders of magnitude higher energy than slow solar wind. On the other hand, it is the solar wind carrying the overall heliospheric magnetic field which is actually limiting the amount of the lower energetic edge of the particles from being able to penetrate into the solar system. The relation does change with the solar activity when overall significantly lower radiation doses are in place during the solar maximum.

Most of the cosmic ray particles create so-called air showers in the Earth's atmosphere, a few kilometres above the ground, by colliding violently with nitrogen and oxygen and creating sprays of secondary particles. This is also the only way to detect the highest energy ones, which are very rare – by reconstructing the original cosmic ray particle that initiated the specific air shower. This is the principle of the Pierre Auger Observatory in Argentina, a high-energy cosmic ray detector, operational since 2008, and run by a large collaboration of scientists from 18 countries.

Some of these particles, like muons, require a significant amount of material to be stopped, so they can only be effectively eliminated in a few laboratories located very deep underground. As we shall see in chapter five, these particles were also theoreticized to play a role in climate change and global warming. As noted earlier, their overall levels are influenced by solar activity. These secondary particles – which you have also just been traversed by while reading this – are also slower. They are the ones that are mainly measured.

The even less energetic particles meet a different fate. After passing through the magnetopause and the various magnetospheric layers, they collide with oxygen and nitrogen atoms at a distance of about one Earth radius, producing neutrons. These neutrons are insensitive to the magnetic field and so they continue from the collision zone in the original direction of the energetic particle, but they are also very unstable. Within a few minutes, they dissociate, creating a proton and an electron. The resulting protons travel back and forth from north to south and south to north along the local line of the Earth's magnetic field until they are eventually absorbed into the Earth's atmosphere. The inner radiation belt is where these particles are stored.

Space Instrumentation and Multi-Satellite Missions

Launching space probes allows both *in-situ* measurements of the properties of the environment as well as the unobstructed observation point above the atmosphere to provide remote sensing of the solar system including the Sun. The atmosphere is transparent mainly to visible light and radio waves, while it absorbs other wavelengths of the electromagnetic spectrum quite efficiently: including infrared, ultraviolet, X-ray and gamma radiation. If the remote sensing processes of interest occur within these wavelengths, it is necessary to observe them from space.

Another reason to be above the lower atmosphere is that it is not completely transparent even to visible light. We can show this by looking at the sunset: at that moment, the sunlight has to pass through a very large mass of air and is strongly attenuated, allowing us to look at the Sun more safely. The light of the stars is not so strong so it is significantly attenuated by the atmosphere even when they are observed at zenith. If we need very precise measurements of the brightness of a star, for example, to follow very subtle variations in brightness, we need to be above the atmosphere.

However, a space mission is expensive, technically difficult, and requires many years of preparation. Consider first the pointing of a telescope in space. There is no solid surface on which the telescope can rest: it fluctuates in space! So some mechanism is needed to keep the satellite stable and pointed in a fixed direction. This mechanism is often a system of rapidly rotating wheels, which maintain the orientation of their axes of rotation in space: this is the principle of the gyroscope.

Think also of the fact that you cannot usually go and repair a satellite if one of its systems fails. Each of its systems must therefore be perfectly reliable. You don't send new, unproven technologies into space!

Satellites are also subject to extreme thermal conditions. The temperature on its side facing the Sun approaches 100 °C, while its side in the shade cools down to −180 °C. Ventilation cooling does not work in space: there is no air! You need other ways to keep your temperature stable and bearable. Yet another problem is the cosmic rays, that we discussed earlier. The atmosphere protects us from most of these very energetic particles, but a satellite is continuously bombarded by them. Without protection, its electronics would quickly be destroyed.

Finally, the often-delicate instruments must survive the vibrations and accelerations of the rocket during launch. Not to mention that not all launches are successful… A failure means the loss of many years of work.

So you need very powerful reasons to build space instruments. Of course, the prestige of a country is measured by its ability to conquer space. But for science, this is not enough… We have already seen how, since the first space missions, we have been forced to totally reconsider our vision of our immediate environment. But crossing regions of different properties with a single satellite quickly leads to ambiguities: did the satellite cross a spatially fixed region or did a moving structure pass 'over' it? To resolve such space–time ambiguities, multi-satellite missions began to emerge in the late 1970s with the International Sun-Earth Explorer (ISEE) 1 and 2 satellites developed jointly by NASA and ESA.

This multi-satellite approach culminated in the 2000s with the fleet of four European CLUSTER satellites, launched in July and August 2000 and de-orbited in September 2024. Before going into the details of the mission and its achievements, the history of CLUSTER deserves a brief introduction. The idea of a multi-satellite mission was born in the minds of Alain Roux (1943–2015) and Michel Blanc in the 1980s with the project proposed to the European Space Agency (ESA). The project was selected in 1985 and was given the name CLUSTER. Ten years of research and development led to the construction of four identical satellites, each equipped with eleven similar instruments, to perform all the useful measurements in the magnetosphere and solar wind. These include electric and magnetic field detectors, mass spectrometers that measure the properties of electrons and ions, wave detectors, etc.

First Unsuccessful Launch of CLUSTER Mission

June 4, 1996, is remembered as a black day for space exploration and astronautics in general. On that day, the maiden flight of Ariane 5 (V501) was supposed to insert the four CLUSTER satellites into orbit from the Kourou space centre. The launch turns into a fiasco, the launcher and its precious cargo are destroyed a few seconds after take-off.

Fig. 3.11 – Ariane launcher 501 exploded on 4 June 1996 with the four CLUSTER satellites on board.

For the women and men, scientists, engineers and technicians, who have been working for more than 10 years on the rocket, the satellite platforms and the instruments, it is the culmination of all their labour and their ambitions that is going up in smoke. It is consternation, sadness and infinite distress that overwhelm all the teams. Along with SoHO (see below), CLUSTER was to be one of the flagship missions, one of the 'cornerstones' of European space exploration, and failure was out of the question. ESA therefore decided to rebuild the mission identically. This time, the launch was entrusted to two Soyuz-Fregat rockets, which launched the mission in two stages, on 16 July and 9 August 2000, placing the four satellites in formation in an elliptical polar orbit with a perigee of 19,000 km (about 3 Earth radii) and an apogee of 119,000 km (about 20 Earth radii).

A tremendous scientific harvest then began. The on-board instrumentation, the chosen orbit and the ability to probe in 3D meant that virtually all regions of the magnetosphere were explored, including key areas of the inner magnetosphere, but also the outer regions of magnetosheath, the bow shock, and solar wind upstream the bow shock. In each region of the Earth's space environment visited, the richness of natural plasma physics and the dynamics of the magnetosphere, which were previously difficult to access, were revealed. Here are some of the most important results, which are by no means exhaustive:

- Upstream of the bow shock, CLUSTER has made it possible to study a whole region called the foreshock, in which the reflection of solar wind particles creates reflected beams and further waves in the medium. This region is important for space weather and space climate because it can significantly modify the properties of the solar wind that are measured upstream and generally assumed to be unchanged until the Earth. But that is valid only for pristine solar wind not magnetically connected to the bow shock.
- Significant advances in our understanding of the penetration of the solar wind into the magnetosphere are being made through CLUSTER. Since the 1980s, observations have suggested that reconnection to the magnetopause is sporadic and impulsive, and occurs as a so-called flux-transfer event. With CLUSTER, we can observe these events in 3D and determine their contribution to the entry of particles and energy into the magnetosphere. It can also be shown that magnetic reconnection does not require rigorously antiparallel magnetic fields to take effect.
- CLUSTER also made the first *in situ* and 3D observations of the reconnection phenomenon in the tail of the magnetosphere with the 'breaking' of the field lines and their subsequent reconfiguration.
- On the flanks of the magnetosphere, a so-called Kelvin-Helmholtz instability can develop due to the difference in speed (known as shear) between the solar wind plasma and the magnetospheric plasma. (In everyday life, this instability is the source of the swell on the sea: if the wind blows hard enough, it causes waves). Thanks to CLUSTER, we realise that within the vortices created, magnetic reconnection can take place.

But perhaps most importantly, CLUSTER takes us from a static, fixed representation of the magnetosphere to a dynamic, ever-changing view. All the beautiful straight lines of the representations in books and on the web are now to be put away in the attic of history, even if they allow a terribly simplified first approach.

In 2023, 23 years after the launch of the CLUSTER mission, it was again prolonged as most of its instruments are still providing data! (Admittedly, some of them are in a degraded form.) Above all, we realise that it will be difficult to return to single-satellite missions to probe an environment as complex as the magnetosphere.

Subsequently, other multi-satellite missions were launched. The first of these, by NASA, was focused on the study of the magnetospheric tail and the triggering of substorms, as suggested by its name – Time History of Events and Macroscale Interactions during Substorms (THEMIS). This five-satellite mission was launched in 2007 and was placed in highly elliptical equatorial orbits with varying apogees to maximise coverage of the extended night side of the magnetosphere. The mission benefits from unique ground support instrumentation as almost all of Canada and the northern United States are covered by networks of magnetometers and full-sky cameras. THEMIS is contributing to a better understanding of substorms even if it does not fully separate the two models (distance reconnection line or current disruption) that we have discussed. In fact, the data seem to favour the reconnection model without excluding the current disruption model.

In 2015, the Magnetospheric Multiscale Mission (MMS) included four satellites dispatched by NASA into (low inclination) equatorial orbit. This mission relies more on small inter-satellite separations and high measurement rates to better understand the regions where the magnetic reconnection process takes place. In particular, it allowed us to observe and understand the so-called diffusion region in which the magnetic fields recombine but also found that this process of magnetic reconnection occurs at unexpected regions so basically all around the magnetosphere.

Ground-Based Instrumentation

Although ground-based measurements are perhaps less prestigious than space-based ones, they provide a unique wealth of knowledge. They have tremendous advantages: they allow for long time scales observations from multiple locations, and instantaneous wide geographic coverage, and some instruments are relatively inexpensive. Let's take a look at these technological marvels that have, like space, changed our conception of the near-Earth environment and allowed space weather to become operational.

The active sounding of the ionosphere started at the beginning of the 20th century with the ionosonde instrument getting well developed including dynasondes

Fig. 3.12 – A dynasonde in Norway, near the Ramsfjørd (credit: M. T. Rietveld, EISCAT Scientific Association).

which vary the emission frequencies automatically, so that different altitudes can be probed continuously. But perhaps most importantly, by taking advantage of two physical processes, coherent and incoherent scattering, optical instruments have also progressed enormously in order to take advantage of the information in optical emissions.

Magnetometer Chains

Historically, the first instruments used to observe and understand the response of the magnetosphere-ionosphere system to solar variability were magnetometers. Simplistically, they could be described as very accurate compasses. In the XVIIIth century, however, scientists had only more or less sensitive compasses with which they measured the direction of the local magnetic field using two angles: magnetic declination – the angle, in the horizontal plane of the location, between the local geographic meridian and the direction of the magnetic field – and magnetic inclination – the angle, measured in the vertical plane containing the direction of the local magnetic field, between the horizontal and the magnetic field. Temporal magnetic variations were therefore only measured in terms of angles. With the progress of electromagnetism and the work of Carl Friedrich Gauss (1777–1855) which he published in 1833, measurements of magnetic intensity became possible. There are many types of magnetometers, the most common of which is based on magnetic induction: a change in magnetic field creates an electric current. Magnetometers working on this principle consist of three metal coils arranged in three different directions. Each of the coils reacts to a component of the magnetic field that is to be measured, and along these coils, an induced current is generated and measured. These instruments are obviously extremely sensitive. Care is taken to install them in isolated locations, far from any source of electromagnetic interference. On satellites, the measurement of the magnetic field is subject to the same constraints. However, the satellite, with all its electronics, is itself a source of interference. This is why space magnetometers are always placed on masts, at a good distance from the satellite body. A magnetometer is an inexpensive instrument that can be installed almost anywhere on Earth and entire networks of these instruments are deployed on all continents: the IMAGE network in Northern Europe, the CARISMA network in Canada, etc. France is a pioneer in this field, and leader of the European INTERMAGNET project, which alone has about 150 magnetometers that meet specific quality requirements. These ground magnetometers have several applications. They give an idea of the geomagnetic activity at a given location in near-real time, as well as an estimate of the direction of the electric current flowing in the ionosphere, which is the cause of the magnetic disturbance measured; grouped together in networks, the magnetometers can be used to reconstruct maps of ionospheric electric currents; and finally, they can be used to calculate indices that give a more global account of the geomagnetic activity (see chapter 6).

Incoherent Scatter Radars

This is a rather anachronistic name for an ionospheric observation technique! The first ionosonde data showed that waves in the MHz range reflect off the ionosphere.

FIG. 3.13 – Far away from everything! Here the French "absolute measurements" shelter at the Dumont d'Urville Magnetic Observatory in Antarctica (credit: Aude Chambodut, Ecole et Observatoire des Sciences de la Terre/Bureau central de Magnétisme Terrestre).

But what happens at higher frequencies? At a few hundred MHz, the ionosphere becomes almost transparent to them. As they pass through, however, they deposit some energy. How does this happen? By causing ions and electrons to oscillate, even though they were already in motion in the atmosphere. Ions are very massive compared to electrons and can be considered fixed. The frequency of oscillation is the same as that of the exciting wave. If we stop giving it energy, this electron regains its initial movement, by simply re-emitting a wave at the same frequency, which will be able to excite other electrons which will in turn start to vibrate. The wave is thus scattered, but the vibration of the electrons starts at a random moment, or in the jargon of radar theory, in an incoherent way. If this were not the case, a wave scattered by an electron could meet a wave scattered by an electron oscillating in the opposite direction (these waves are said to be in phase opposition): they would destroy each other.

Obviously, many things can change the frequency of the scattering wave. If the ionospheric environment is dense, collisions will accelerate or slow down the electrons, which will therefore emit waves around the emission frequency. Temperature is also an important factor. The electrons can never move very far away from the ions, so the ions, even though they are hardly set in motion by the emitted wave, play a role. The initial movement of the ions and electrons in relation to the observer also produces a change in the frequency observed by the Doppler effect. Finally, it is sufficient to measure the time between emission and reception to know at what altitude the wave was backscattered.

Based on this relatively simple principle, one could imagine that by sending a wave into a plasma and then 'listening' to its deformation after scattering, one could deduce at all ionospheric altitudes – typically from 70 to 600 km – the ion and electron number densities, their temperatures and their overall speed! A lot of parameters for a single instrument! But how can such a feat be achieved? The wave emitted by the radar goes up into the sky, but is backscattered in all directions: what comes back to the radar is such a small portion that it seems impossible to measure. Unless you have very large antennas…

The implementation was pioneered by the British scientist Gordon in 1958, taking advantage of the increasing sensitivity of detectors. Bowles then made the first detection, paving the way for an impressive series of instruments built mainly by the United States: from 1962 at Jicamarca in Peru, with 18 000 small antenna – dipoles – spread over a square of 300 m, then at Millstone Hill in Massachusetts in 1963, with a parabolic antenna 70 m in diameter, before creating a veritable emperor at Arecibo (Puerto Rico) in 1964, with what has remained the largest antenna in the world – 305 m in diameter – until 2018 when it got heavily damaged by a hurricane and collapsed. In 1979, they commissioned a new incoherent scatter radar in Poker Flat, Alaska, near a small rocket launch base: they could send instruments into the ionosphere and measure its parameters with the radar!

In France, researchers (Bauer, Petit, Giraud, then Kofman, Blanc, Alcaydé…) had a wonderful idea. To see in three dimensions, we need two eyes. Similarly, if we observe a volume of the ionosphere from two points, we should be able to deduce the direction of the motions, the "vectors". Thanks to the French government's policy of major works, in 1965 they built a transceiver at Nançay (in the centre of France) and an additional receiver at Saint-Santin (further South). The results lived up to expectations. For the first time, the direction of the ions in the upper atmosphere could be distinguished! But the latitude is not the best. As we have seen in the previous chapters, the dynamics of the upper atmosphere are linked to the geomagnetic field. To understand physics, it is easier to place yourself perpendicular to this field, *i.e.* near the equator, like Arecibo or Jicamarca, or parallel to it, *i.e.* at high latitudes, like Poker Flat in Alaska. A group of European countries, Germany, France, Norway, Sweden, Finland and Great Britain, capitalised on the valuable experience of the French and in 1981 opened the flagship of incoherent scatter: the scientific association EISCAT and its four antennas, including two transmitters at Ramsjførden, 30 km from Tromsø in Norway, well north of the arctic circle.

In the Soviet Union, the first efforts date back to 1964, under the influence of a young researcher, Geliy Zherebtsov (born in 1938), but it was not until 1981 that the first Soviet incoherent scatter radar began observing the ionosphere in Siberia, followed by another in Irkutsk (at mid-latitude) in the late 1990s.

Fig. 3.14 – EISCAT Svalbard Radar (ESR) on Spitzbergen (credit: J. Lilensten/CNRS).

China opened the FAST radar in 2016 in the Guizhou province. Nicknamed the ‘eye of the sky’, it is very similar in design to Arecibo, but would certainly be its very big brother, with its 500 m diameter antenna, which makes it the largest astronomical instrument in the world today (the Russian Ratan 600 radio telescope consists of separate radio antennae arranged on a circle 576 m in diameter). This is all the more important as Arecibo was shut down after several hurricanes and a cable break in 2020 and needs a serious facelift. FAST’s mission is to discover the laws of evolution in the universe, so it does not have a transmitter that can be used to implement incoherent scattering. However, China is pursuing a major project, both civilian and military, with a chain of several incoherent scatter radars whose characteristics and objectives are still unclear.

Incoherent scatter radars are the Rolls Royce of space weather. But they are so expensive that it is impossible to achieve global coverage of the Earth. What remains are essentially research instruments, which are indispensable for space weather.

EISCAT 3D

In Northern Europe, the replacement of incoherent scatter radars is ensured with a new generation system that allows for unprecedented temporal and spatial resolutions, while maintaining the ability to probe in three dimensions. This is EISCAT 3D. Instead of using heavy parabolas, an EISCAT 3D radar is made up of nearly 10,000 small antennas grouped by 91 on 109 hexagonal modules! By combining the transmission and reception of these antennas, using the technique of aperture synthesis, a beam is formed in the desired direction. The advantage is that you can change direction much more quickly than with a parabola and thus probe an entire volume of the ionosphere in a very short time, typically one second (compared with several minutes with the old system). The system initially comprises a transmitting-receiving site at Skibotn in Norway and two receiving side stations at Karesuvanto in Finland and Kaiseniemi in Sweden. In the second phase, it will receive two more stations.

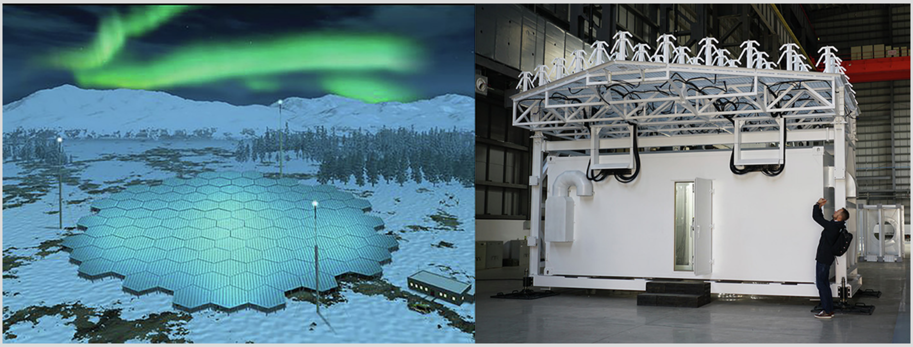


Fig. 3.15 – Left: Artist's view of an EISCAT 3D radar with its 109 modules (credit: NIPR). Right: A module with 109 small antennas on its electronics and computing container (credit: Craig Heinselman).

Coherent Scatter Radars

Incoherent scatter radars are wonderful instruments which allow a very complete diagnosis of ionospheric thermodynamics, but they have two major defects: their range is relatively limited (of the order of 1,000 km at most), even with transmitters of several GW; and the power of the backscattered signal is so low that the antenna has to be left in the same position for several tens of seconds (typically 1 min for EISCAT) to accumulate an exploitable signal. This time is too long when one wants to scan a whole area for example. In order to have a larger-scale view of ionospheric (and therefore magnetospheric) dynamics, another category of radar, operating on a slightly different physical process, was developed at the end of the 1970s: coherent scatter radars.

The principle of the radar is still the same: a wave is sent into the medium and the backscattered wave, whose properties have changed slightly, is picked up. From these changes, information is obtained about the medium being scanned. In the case of coherent scatter radars, waves are emitted at lower frequencies, a few tens of MHz (high-frequency range, or HF), and the scattering process – known as Bragg scattering – is as follows: the emitted wave is backscattered by irregularities in the density of the plasma naturally present in the ionosphere. There is one condition, however: the emitted wave must arrive perpendicular to the magnetic field lines to be backscattered. Since at high latitudes, the geomagnetic field lines are almost vertical, HF waves must be transmitted almost horizontally. From the received signal, the main parameter inferred is the horizontal velocity of the ionospheric plasma. Coherent scatter radars provide less information than incoherent scatter radars, but they have a much greater range of about 3,000 km and require less observation time to pick up sufficient signals. This means that a very large horizontal extent of the ionosphere can be scanned in a relatively short time, of about two minutes.

FIG. 3.16 – The SuperDarn radar in Stokkseyri.

After a few precursor radars in the 1970s and 1980s, such as the STARE radar in northern Europe, the SuperDARN coherent scatter radar network was created in the mid-1990s. In 2020, it comprised 35 radars covering the polar ionospheres of the northern and southern hemispheres. Most of the radars are paired so that the same geographical area is covered by at least two radars, allowing the value and direction of the plasma velocity to be reconstructed. With the whole network, it is possible to map the horizontal motion of the ionospheric plasma (the 'convection') in near-real time.

Measuring the Total Electron Content

This is a technique that came about by opportunity. In the 1980s, the United States made its global positioning system, GPS, public, followed by the Russian (Glonass), Chinese (Beidou), Sino-Indian (IRNSS) and European (Galileo) systems. They all work on the same principle. A constellation of satellites (between 24 and 30) flies at altitudes of about 24 000 km so that any observer on Earth can see at least four of them at any given time. Each satellite gives the following information: "I am satellite number ***, I am at such and such a place and it is such and such a time". The ground system compares the time of arrival of the information with the time of transmission and deduces how far away it is from the four satellites. From this, it is easy, with a little geometry (called trilateration), to deduce its own position.

One of the uncertainties in the measurement is the number of electrons that the wave encounters between the satellite and the receiver. These electrons vary the path of the wave, like a prism for visible light. You thought the wave had travelled 40 000 km in a straight line? In reality, it travelled 42,000 km in a zigzag pattern, resulting in a positioning error of several metres vertically and several dozen centimetres horizontally. To counter this problem, geodetic stations have been built all around the Earth, whose position is known with a precision of the order of a millimetre. The difference between this actual position and the one indicated by the system allows corrections to be made elsewhere in the world. But for space meteorologists, it does something else: it calculates the total electron content (TEC) encountered by the wave on its way from the satellite to the receiver. By pooling data from a large enough number of ground-based receivers, global TEC maps can be reconstructed that give an idea of the ionisation of the Earth's upper atmosphere. This is a by-product of these satellite systems that was certainly not expected in the 1980s, and which has given them some added value.

Optical Instruments

How can one imagine working on the aurora without observing its visible light? Using telescopes is not relevant: they are designed to explore bright, point-like objects, whereas auroras are diffuse and extensive. You might as well use fast cameras! For better coverage, some of them are equipped with a very wide angle, a fish eye, which allows them to observe the entire sky. They are then nicely called all-sky cameras.

Other variations exist. Some, with a restricted angle of view, scan the sky from north to south with a motorised mount. Others take into account the fact that the most intense emissions are well known – mainly red, green and purple – and that it is useless to waste time observing colours that will never be emitted by an aurora. A camera that observes only one colour can even count the photons it receives: it becomes a photometer.

Fig. 3.17 – In the Arctic before a snowstorm, protection of an auroral photopolarimeter (credit: J. Lilensten/CNRS).

We can try to count the photons in all the measured wavelengths. This is called a spectrometer, or a hyperspectral camera.

There are many models, but the basic principle remains the same. As with all optical instruments, they are limited in their use: you need a clear, moonless sky and a carefully cleaned astronomical dome. But their usefulness is undeniable.

Other instruments are more sophisticated. The red auroral line, for example, is emitted at 630 nm. But although the oxygen that causes it has its own movement, it is received with a very slight shift – a few hundredths of a nanometre – due to the famous Doppler effect. Such a very small deviation can nevertheless be measured using an interferometer, which therefore gives access to the wind speeds of the upper atmosphere.

Neutron Monitors

As we have seen, some of the cosmic rays are stopped by collision as they pass through the Earth and some of them create secondary particle sprays in the atmosphere. How can we measure them? The principle is relatively simple. Let's take a watertight box. Fill it with a thick, conductive liquid, such as leaded liquid. This makes the box considerably heavier, of course, up to several tonnes. This is a difficult crossing for cosmic particles which will interact in it producing a dozen neutrons. All that remains to be done is to slow them down and then count them. To do this, they are passed through a material where they produce an electrical signal.

The lower the altitude, the faster the blocked particles arrive. It is therefore tempting to measure at sites several hundred metres deep and thus be able to measure part of the highest energy spectrum of cosmic rays. From this measurement,

FIG. 3.18 – Professor Chilingarian in front of the cosmic ray air shower particle monitoring cabinet under Mount Aragats, Armenia (credit: J. Lilensten/CNRS).

and with a little physics, it is even possible to extrapolate to other energies. And if you measure over large areas rather than a column, you can even assess where the cosmic particles come from.

There are several such neutron monitor farms around the world. During the Second World War, the Germans used their prisoners to dig a monstrous cave in the heart of Mount Aragats in Armenia. The Soviets continued this cyclopean work by using their German prisoners. The aim was, for both sides, to make a weapons storage. However, by the time the project was completed, the original purpose was obsolete. A researcher, Ashot Chilingarian, used the underground temple as the world's largest neutron monitor farm with many other cosmic ray air shower particle monitoring instruments like muon telescopes.

References

Akasofu S.-I. (2015) Paradigm transitions in solar–terrestrial physics from 1900: my personal view, *Hist. Geo Space. Sci.* **6**, 23–43, https://doi.org/10.5194/hgss-6-23-2015.

Ammar-Israël A. (Ed.) (2015) 50 ans de coopération spatiale France-URSS/Russie, Genèse et évolution 1966–2016, Tessier & Ashpool.

Bauer P., Giraud A., Kofman W., Petit M., Waldteufel P. (2013) How the Saint Santin incoherent scatter system paved the way for a French involvement in EISCAT, *Hist. Geo Space Sci.* **4**, 97–103, www.hist-geo-space-sci.net/4/97/2013/, https://doi.org/10.5194/hgss-4-97.

Feldstein Y. I. (1986) A quarter of a century with the auroral oval, *Eos Trans. AGU* **67**(40), 761–767, https://doi.org/10.1029/EO067i040p00761-02.

Gordon W. (1958) Incoherent scattering of radio waves by free electrons with applications to space exploration by radar, *Proc. IRE* **46**(11), 1824–1829. https://doi.org/10.1109/JRPROC.1958.286852.

Gubbins D., Herrero-Bervena E., (Ed) (2007) Encyclopledia of geomagnetism and paleomagnetism, *Encycl. Earth Sci. Ser.* Springer.

Lemaire J. (2002) La plasmasphère, *Physicalia Mag.* **24**(4), 231–252.

Lockwood M. (2016) Jim Dungey, the open magnetosphere, and space weather, *Space Weather* **14**, 380–383, https://doi.org/10.1002/2016SW001438.

Risbeth H. (2001) The centenary of solar-terrestrial physics, *J. Atmos. Sol.-Terr. Phy.* **63**, 1883–1890.

Stern D. P. (1996) A brief history of magnetospheric physics during the space age, *Rev. Geophys.* **34**, 1.

Van Allen J. A. (1959) The geomagnetically trapped corpuscular radiation, *J. Geophys. Res.* **64** (11), 1683–1689, https://doi.org/10.1029/JZ064i011p01683.

Van Allen J. A. (1983) *Origins of magnetospheric physics*, University of Iowa Press.

Chapter 4

The Time of Complexity: The Sun

At the start of the space age, during the 1950s and the 1960s, the launch of satellites allowed scientists to advance significantly in our understanding of the ionosphere, thermosphere and magnetosphere, due to our ability to probe these environments *in situ* for the first time in history. In 1960, we observed Earth from space for the first time, allowing us to image weather patterns. It was only a few years later, in 1962, that the first space telescope was launched and set out to observe our Sun. A lot of advantages come with sending a spacecraft up in space to observe our Sun. By removing the influence of Earth's atmosphere, we are able to observe the full spectrum of the Sun, without parts being absorbed by Earth's atmosphere and without being limited by turbulence or visibility in case of clouds. From a science perspective, the solar wind and the Sun-Earth environment can be probed *in situ* and more continuously. It is under these advances that our knowledge of the Sun and the Sun-Earth environment expanded immensely.

The Dynamic Sun and the Solar Wind

In the 1940s, Grotain, Edlén, and Alfvén showed that the solar corona is likely a million degrees hot. Since the temperature of a body determines its spectral emission, the spectrum of the solar corona should peak at a wavelength of about 30 Å, *i.e.*, not in the visible light but close to the extreme ultraviolet and soft X-rays. However, at these wavelengths, Earth's atmosphere is opaque; if we want to observe the solar corona at its hottest temperatures, we have to go into space. As we learned already in previous chapters, building on the idea of a hot corona, Eugene Parker concluded in the 1950s that the Sun must be emitting a continuous stream of particles. To prove this theory, *in-situ* observations in interplanetary space were needed. Fortunately, the space race started right around this time, allowing scientists to probe the solar wind *in situ* and observe the Sun in X-rays.

The first direct measurements of the solar wind were made by the Soviet Luna 1 spacecraft on its route to the moon in 1959 and by the US Mariner 2 spacecraft on its route to Venus in 1962. Both found a continuous stream of charged particles travelling away from the Sun, confirming the solar wind theory. The solar wind speed was not constant but varied between 300 and 850 km/s. The slow portion of

DOI: 10.1051/978-2-7598-3610-9.c004

the solar wind, usually defined by speeds less than 400 km/s, is called the slow solar wind. All faster solar wind streams were called high-speed solar wind streams.

Between 1962 and 1975, the USA launched a series of 8 space telescopes within the Orbiting Solar Observatory (OSO) programme. These telescopes were mainly dedicated to studying the Sun, and they included UV and X-ray telescopes for the first time. These telescopes were still in their infancy but could already show that the Sun indeed radiates in X-rays, *i.e.*, that the solar corona is indeed a million degrees hot, and that the solar corona is not uniform but contains large-scale structures. Dark regions and exceptionally bright regions were observed for the first time. The dark regions were named coronal holes, while the bright regions coincided with the locations of sunspots in the photosphere and were thus suspected to be their coronal counterparts. Today, the bright regions are referred to as active regions.

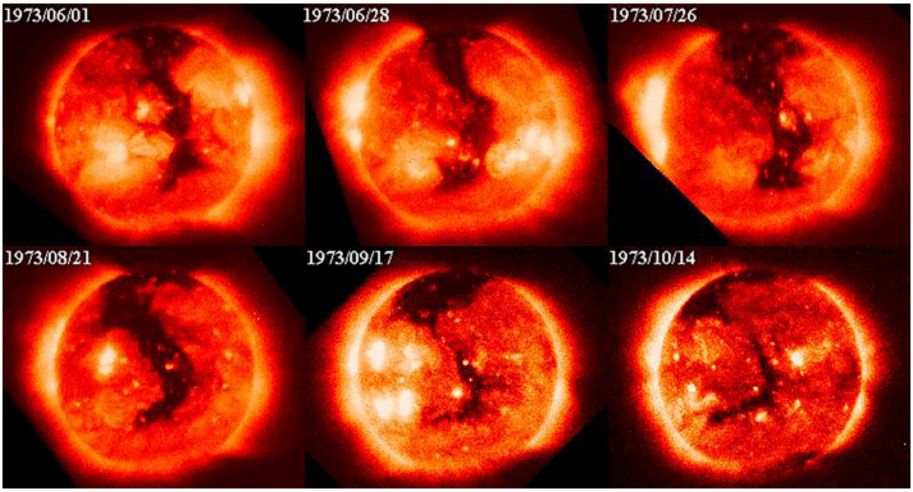


FIG. 4.1 – The Sun in X-rays observed from Skylab. The black areas in the images above are the coronal holes (credit NASA).

Due to these results, the USA launched the space station Skylab in 1973. Skylab hosted the Apollo Telescope Mount (ATM) solar observatory, which included several instruments: two X-ray telescopes, an ultraviolet spectroheliograph, an ultraviolet spectroheliometer and spectrograph, a visible coronagraph, and two Hα telescopes. Astronauts aboard the Skylab space station could study the corona with the ATM for a total period of 9 months, which, considering the era during which it was launched, is quite significant. They obtained unprecedented results and produced some images that are, even today, still remembered by the public.

Skylab was able to scan the solar corona continuously for coronal holes, and scientists were able to study their evolution and analyse their rotation in and out of the visible side of the solar disk. By comparing the appearance of coronal holes on the Sun with *in-situ* solar wind measurements near Earth by the Interplanetary

Monitoring Platform (IMP) 7 and 8, and with magnetic field measurements at the Earth's surface, it became clear that each time a coronal hole appeared close to the centre of the solar disk, *i.e.*, facing Earth's direction, about two to six days later a high-speed solar wind stream was measured by IMP close to Earth. Furthermore, when a high-speed stream was observed, geomagnetic disturbances were also recorded by the magnetic ground stations on Earth. Therefore, Bartel's solar M-regions, which he thought were responsible for the recurrent geomagnetic storms, were finally identified to be coronal holes. These observations finally proved that a solar-terrestrial relationship exists.

Apart from coronal holes, Skylab was able to record many other solar phenomena, such as solar flares, *i.e.*, a sudden burst releasing energy in all wavelengths, from radio to gamma rays, and coronal mass ejections, *i.e.*, large-scale eruptions of magnetised plasma that is ejected from the Sun at large speeds. The astronauts aboard the Skylab space station managed to take detailed recordings of 8 major flares – which was an amazing achievement considering that solar flares are fast-evolving events which typically happen within the timescale of tens of minutes, and that the astronauts had to manually start the dedicated observation at this time. Furthermore, Skylab registered more than 400 smaller flares and more than 100 coronal mass ejections by routine observations.

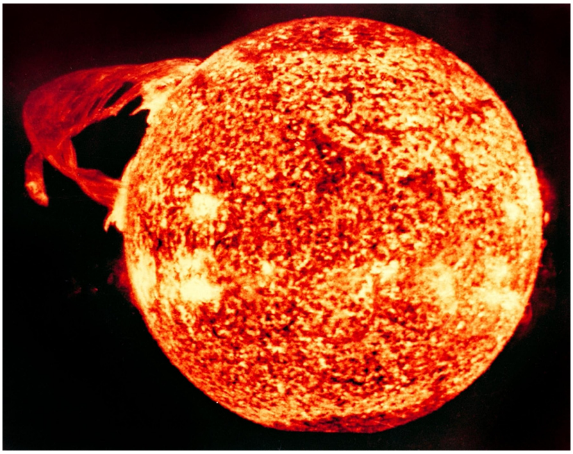


FIG. 4.2 – Extreme ultraviolet image of a coronal mass ejection taken aboard Skylab (credit NASA).

The Solar Flare Myth

Following Carrington's flare observations in 1859, which were associated with auroras that were visible up to low-latitudinal regions such as Rome and Hawaii, a statistical study by H. W. Newton on the relationship between solar flares observed in Hα and intense geomagnetic storms in 1943, and the new X-ray flare observations by Skylab, the idea that solar flares are the source of intense geomagnetic storms became a paradigm. It was thought that large solar flares could emit streams of plasma into interplanetary space, which cause non-periodic, intense geomagnetic storms when they hit Earth.

Flares seemed to be important, and consequently, the Solar Maximum Mission (SMM) satellite was designed. It aimed to study the active Sun and particularly solar flares together with their terrestrial response in more detail. SMM was launched by NASA in 1980, and taking advantage of the newly inaugurated Space Shuttle, SMM was designed to be repairable. This was a fortunate decision as a fuse in the pointing system burned out just nine months after launch, leaving the satellite unable to point precisely at observation targets on the Sun. Fortunately, it was repaired by the crew of the Space Shuttle Challenger in 1984 and continued its mission until 1989. During its lifetime, it observed more than 12,000 flares and more than 1200 coronal mass ejections. The flare paradigm seemed to hold: the majority of the intense geomagnetic storms were preceded by a solar flare.

However, there was one issue: *in-situ* measurements of the solar wind near Earth showed that all intense geomagnetic storms were accompanied by a shock wave and a strong southward-oriented component of the interplanetary magnetic field, but the observed flares did not produce a visible shockwave in solar imagery and also could not explain the southward-oriented magnetic field component.

How to solve this discrepancy? There was another solar phenomenon, the coronal mass ejection, which seemed to produce a shockwave in solar imagery. Coronal mass ejections result from unstable magnetic field configurations and are thus related to solar flares and eruptive prominences. The magnetic field topology of coronal mass ejections was able to explain the prolonged southward-oriented magnetic field component that was being measured near Earth. Furthermore, for each of the intense geomagnetic storms, a coronal mass ejection on the Sun had also been observed a few days earlier. Therefore, the correlation between the co-occurrence of flares and intense geomagnetic storms is not a direct causal relationship – instead, flares can launch coronal mass ejections which then cause geomagnetic storms. This was described by Gosling in 1993 in an article named 'The Solar Flare Myth', which caused a change of paradigm in the view of the solar sources of the intense geomagnetic storms on Earth.

The Solar Dynamo and Solar Cycle

Having set up the solar players for space weather and space climate – coronal holes, flares, and coronal mass ejections – scientists aimed to understand their physics. To that end, the Solar and Heliospheric Observatory (SOHO) satellite and its successor, the Solar Dynamics Observatory (SDO), were designed. In contrast to the preceding missions, SOHO and SDO were monitoring missions, meaning that many of their instruments took data at regular intervals. SOHO was launched in 1995 and placed in orbit around the 1st Lagrangian point, a point 1.5 million km from the Earth in the direction of the Sun. In this location, the gravitational forces of the Sun and Earth balance each other and SOHO can remain orbiting around the line between the Earth and the Sun, providing the spacecraft with an uninterrupted view of the Sun. SOHO provided extreme ultraviolet (EUV) images of the Sun in four wavelength bands, which correspond to four different coronal temperatures, at a cadence of 12 minutes, as well as coronagraphic images at a cadence of 20 to 30 minutes, and magnetograms and dopplergrams at a cadence of 96 minutes. The mission was supposed to last three years. In 1998, once the nominal mission was completed, it was extended, given the quality of the scientific results. Since then, the mission has been regularly extended, and some of the instruments are still operational today. SDO was launched in 2010 for a five-year mission with an expectancy of ten years. SDO is also still operational today and is one of the main data sources for solar scientists nowadays. SDO takes images in the continuum every hour, images in two UV windows every 24 seconds, images in seven EUV wavelength bands every 12 seconds, and magnetograms and dopplergrams every 45 seconds, all at a 4 k resolution. By that, SDO produces about 2 terabytes of solar data each day. To be able to transmit this huge amount of data back to Earth, SDO was put in a near-Earth orbit at an altitude of 36 000 km.

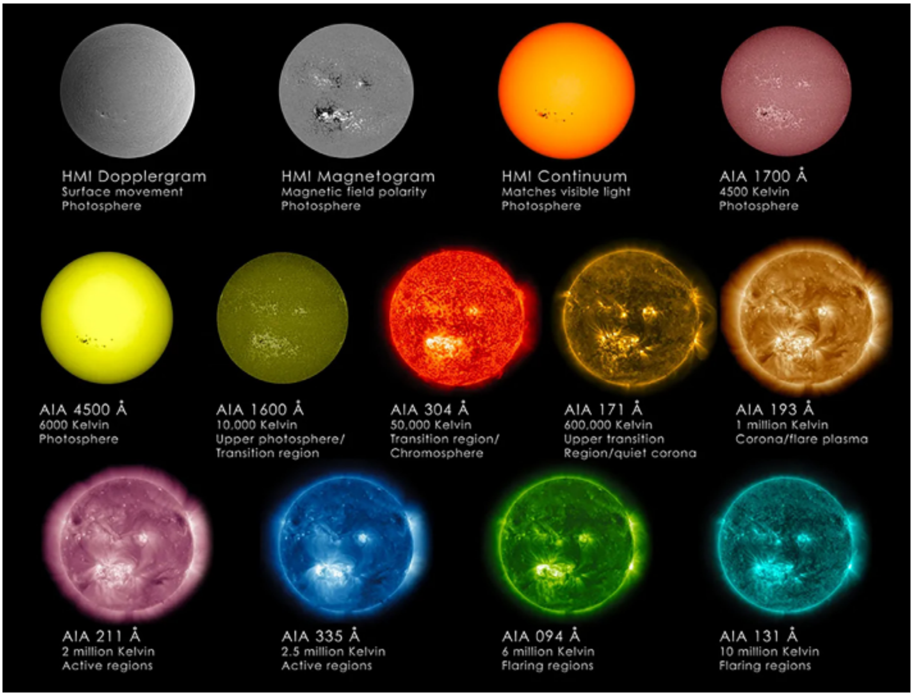


FIG. 4.3 – The Sun as observed by SDO (credit NASA, C. Alex Young).

Why do the satellites take images in so many spectral bands, and particularly, why in the extreme ultraviolet? The images in the extreme ultraviolet waveband correspond to highly ionised states of chemical elements such as iron, which are a result of the high temperature of the corona. Each element emits certain spectral lines depending on its temperature. Therefore, observing a certain spectral band enables one to see the solar corona at an associated temperature level, *i.e.*, at an associated atmospheric layer. The Extreme Ultraviolet Imaging Telescope (EIT) aboard SOHO took images of the Sun at wavelength bands of 17.1, 19.5, 28.4 and 30.4 nm, which correspond to solar atmospheric temperatures of 900,000 K, 1.3 million K, 2 million K, and 80,000 K. The Atmospheric Imaging Assembly (AIA) instrument aboard SDO has a spatial resolution four times higher than SOHO's EIT and observes the solar atmosphere in seven EUV channels: 9.4, 13.1, 17.1, 19.3, 21.1, 30.4, and 33.5 nm. These wavebands span the temperature range from 50,000 K to 20 million K as described in the picture.

To be able to distinguish the images in the different wavelengths, which are in a range of the electromagnetic spectrum that is invisible to the human eye, scientists agreed to associate specific colours with the wavelengths. In visible light, the colour red corresponds to the electromagnetic spectrum emitted by cool bodies, and the colour blue to the spectrum of the hottest bodies. Following this same logic, the images taken by EIT and AIA were also assigned a colour: red for the channels probing cool solar plasma, and blue for the channels probing hot solar plasma.

As SOHO and SDO captured images from the Sun in EUV for over 25 years, both spacecraft provided scientists with enough data to learn more about the dynamics of the corona and the phases of the solar cycle. The continuous observations clearly showed a corona dominated by a dipolar (magnet-like) magnetic field at solar minimum, and a much more complex corona during maximum activity. In the adjacent series of images taken by EIT over 20 years, the corona is dark and quiet looking in 1995–1996 and again in 2009–2010, *i.e.*, in the solar minima, but bright and dishevelled in 2000–2001 and again in 2014–2015, *i.e.*, in the solar maxima.

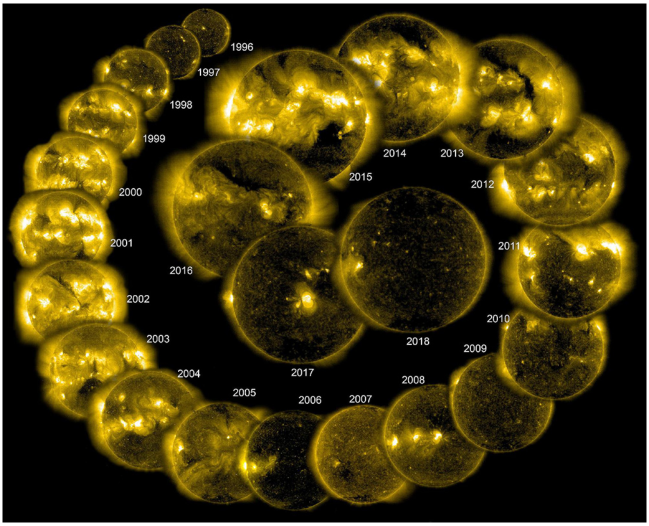


FIG. 4.4 – SOHO/EIT images of the Sun over 20 years.

We Lost SOHO!

The lifetime of a satellite is not always a long and quiet river. On 24 June 1998, there was a problem: radio silence… We lost SOHO!

SOHO has three gyroscopes to help it maintain its attitude. Gyroscopes need to be calibrated regularly during routine operations. On this specific day, this routine operation went wrong. One step of the procedure was not carried out and the gyroscopes stopped providing the correct orientation without the controllers noticing. When they commanded SOHO to point towards the Sun (using its booster nozzles), the satellite's attitude control was lost. With the communication antennas pointing in the wrong direction, the controllers could no longer communicate with SOHO: the satellite, a billion-euro mission, was lost in space… Over the next month, all communication attempts failed: SOHO no longer responded, and no one knew where it was.

FIG. 4.5 – The SOHO satellite in its assembly hall.

The University of Colorado proposed using the Arecibo radio observatory, back then the largest antenna in the world (with a diameter of 305 m). The idea was to use this antenna as a radar, but it could not transmit and receive at the same time. As such, NASA decided to use the Arecibo antenna as a transmitter, and one of the antennas of its Deep Space Network (DSN) as a receiver. On 23 July, the Arecibo antenna transmitted a signal to the suspected location of SOHO. The DSN antenna received the signal reflected from SOHO and so the satellite was finally found: it remained close to its predicted location, slowly rotating.

Communication was slowly restored. First, the batteries had to be charged by the solar panels; on 8 August enough power was available to activate the telemetry. The first news was not very good: the fuel was frozen. Once the batteries were fully recharged, the fuel tank was gently warmed up, a process that took more than 10 days.

In the end, SOHO was able to point at the Sun again. The instruments were reactivated and tested. Surprisingly, almost all of them still worked perfectly, except for a single coronographic camera which was lost for good. On 24 October 1998, SOHO resumed its work after a four-month holiday. The recovery of SOHO remains one of the most spectacular rescues in space!

In addition to these amazing coronal observations, the long-time series of Dopplergrams provided by these satellites gave a huge boost to the field of helioseismology. Helioseismology is the analogous field to seismology on Earth. By analysing long time series of oscillation patterns on the solar surface, scientists could derive sound speeds in the solar interior and infer properties like density, temperature, magnetic field profile, as well as the solar rotation period in the solar interior. SOHO data showed that theoretical models of the Sun's interior as well as the solar dynamo were quite correct. In the Sun, plasma and magnetic field are frozen-in, meaning that plasma and magnetic field can only move together. Because the solar rotation rate is faster at the solar equator than at the solar poles, the solar dipole field in the solar interior gets stretched, slowly winds up, and thereby increases the magnetic field strength in the solar interior. A magnetic field always involves magnetic pressure, which is the electromagnetic analogue to the gas pressure. When the magnetic pressure is sufficiently high, the magnetic field and the frozen-in plasma expand, rise to the solar surface and, due to the Coriolis force, rotate in a way that the magnetic field is oriented in the opposite direction compared to the original solar dipole field. This rotation of the magnetic field direction together with the transport of magnetic flux from the equator to the poles due to latitudinal plasma flows results in a polarity reversal of the Sun's dipole field roughly every 11 years. This process defines the solar cycle. A full magnetic solar cycle hence takes 22 years, as two reversals are needed to create the original polarity again.

By linking the helioseismic information on the solar dynamo with the appearance of sunspots in the photosphere and the features in the solar corona, we can finally explain the appearance of sunspots and active regions. The locations where

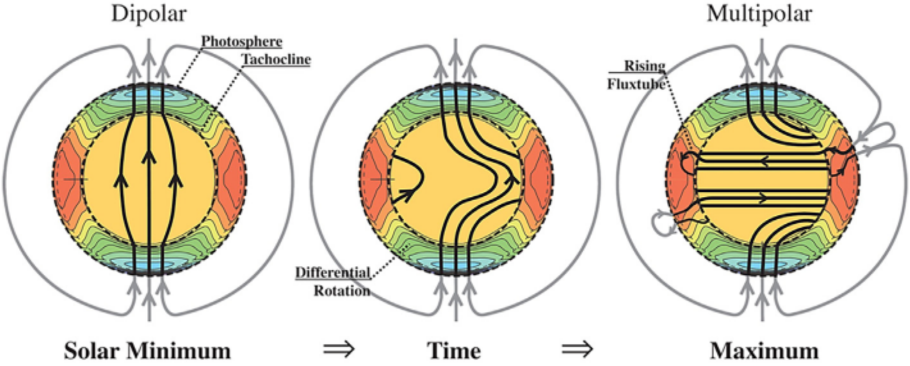


FIG. 4.6 – The solar dynamo. The solar differential rotation winds up the magnetic field, enhancing the magnetic field strength. The magnetic field rises, rotates and pierces through the surface, and becomes visible as sunspots. The magnetic field piercing through the surface is oriented in the opposite direction compared to the dipole field and eventually results in a polarity reversal.

the ascending magnetic field of the solar dynamo pierces through the solar surface are visible as a pair of sunspots in the photosphere, one sunspot where the magnetic field exits from the solar interior through the photosphere and one where the magnetic field re-enters into the solar interior from the photosphere. Therefore, one of the spots has a negative magnetic polarity and one a positive magnetic polarity. Between the two spots, a magnetic loop rises into the solar corona, which forms the active region. The stronger the magnetic field in the solar interior, the more sunspots appear on the solar surface. When two active regions, one in each solar hemisphere, decay, the magnetic flux from the equatorward spot of each active region diffuses, partially wanders to the equator, and cancels. The magnetic flux from the remaining spots diffuses and is slowly transported towards the solar poles. During that transport process, large-scale regions with a dominant magnetic polarity, *i.e.*, the polarity from the remaining spot, may form. These regions are predestined to have coronal holes. When the magnetic flux in these regions is sufficiently large, it might happen that the magnetic field does not close near the Sun anymore, but reaches far into interplanetary space. Along these open-to-interplanetary-space magnetic field lines, solar plasma is accelerated to form high-speed solar wind streams. The outflowing solar wind transports mass and energy away from the Sun, which results in a locally reduced coronal density and temperature. The reduced density and temperature result again in a reduced emission, which is why coronal holes appear dark in the EUV images. Since the appearance of coronal holes is linked to the magnetic flux distribution of the Sun, coronal holes are another expression of the solar dynamo process. During the solar maximum, when many sunspots are present, coronal holes appear at all latitudes. During the solar minimum, when the sunspots have vanished, large coronal holes predominantly form around the solar poles.

The Solar Wind

Having understood active regions and coronal holes in the solar corona, we next zoom out from the Sun and target the associated solar wind structure in interplanetary space. In 1962, when Mariner 2 took its solar wind measurements on its route to Venus, it became clear that the solar wind is built up of two components, high-speed solar wind streams and the slow solar wind regime. Skylab clearly showed that high-speed streams originate from coronal holes, but the solar source regions of the slow solar wind are still debated even nowadays. Since then, a variety of satellite missions dedicated to understanding the origin, properties, and three-dimensional distribution of the solar wind in the heliosphere have been launched. The most prominent ones are the IMP satellites between 1963 and 1973, the Helios satellites in 1974, the Voyager satellites in 1977, the Ulysses spacecraft in 1990, the Wind spacecraft in 1994, the Advanced Composition Explorer (ACE) in 1997, the Parker Solar Probe in 2018, and the Solar Orbiter in 2020.

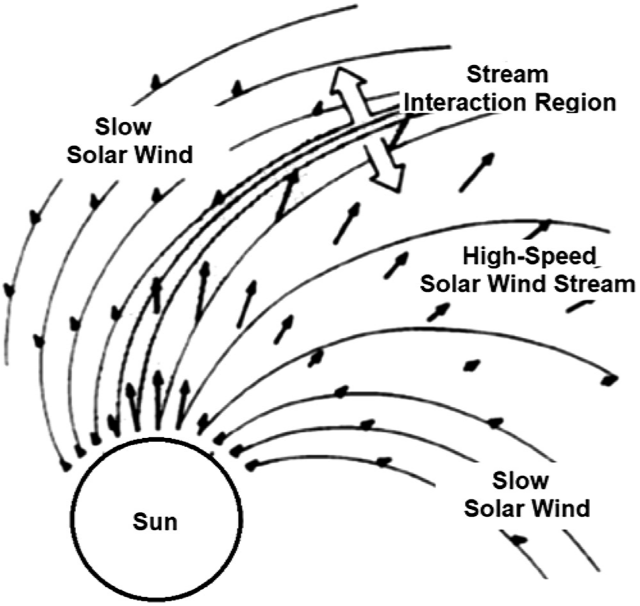


FIG. 4.7 – Schematics: interaction between the ambient slow solar wind and the subsequent rarified high-speed solar wind stream. The arrows give the solar wind velocity, the lines correspond to the magnetic field orientated along the Parker spiral.

A general picture of the solar wind was already formed in the 1970s by combining Eugene Parker's theory of the solar wind with *in-situ* solar wind measurements of Mariner 2 and the IMP satellites, and with solar remote observations by Skylab. Let us first start at the solar corona. Since the Skylab era, it became clear that the corona is almost entirely determined by the magnetic field, and plasma can only move along the magnetic field lines. In coronal holes, the magnetic field lines are open to interplanetary space. Consequently, plasma from coronal holes can leave the corona and be accelerated towards interplanetary space, forming the high-speed solar wind streams. High-speed streams propagate radially away from the Sun and have velocities up to 800 km/s near Earth. This indicates that due to their high speed, they can reach Earth in only two days!

As predicted by Parker, during the propagation phase of high-speed streams from the Sun to Earth, the magnetic field of high-speed streams stays linked with their source coronal holes on the Sun. The Sun rotates, and so do the coronal holes at its surface and consequently the direction in which the high-speed stream plasma from the coronal holes is launched into interplanetary space. This results in a spiral-like solar wind structure in interplanetary space called the Parker spiral. The magnetic field, frozen in the solar wind plasma, looks like it winds up and is oriented along the Parker spiral. At Earth, the angle in the ecliptic between the magnetic field and the Sun-Earth line is typically 45–60 degrees.

The slow solar wind propagates in a similar way through interplanetary space as high-speed solar wind streams. The slow solar wind has a typical velocity of 200–400 km/s, and its chemical composition is highly variable. Also, the slow solar wind propagates radially away from the rotating Sun and its magnetic field also forms a sector of the Parker spiral.

As high-speed streams are faster than the slow solar wind, high-speed streams eventually catch up with preceding slow solar wind plasma and form a stream interaction region. At this stream interaction region, high-speed stream plasma accumulates at the back of the stream interface and slow solar wind plasma piles up in front of the stream interface, compressing the plasma in the stream interaction region to a strongly enhanced density, temperature, and magnetic field strength. This interaction region is usually the geoeffective part of the high-speed stream that causes the geomagnetic storms. High-speed streams typically do not produce the strongest geomagnetic storms, but they appear very frequently and their strength is still significant.

Due to its geoeffectiveness, the solar wind has been monitored near Earth since 1963, first by the IMP satellites, and nowadays by Wind, ACE, and DSCOVR. However, measurements near Earth could not satisfy the curiosity of scientists, and quickly, the question arose how the properties of the solar wind change throughout the solar system. First, they aimed at the inner solar system. In 1974, the twin Helios satellites were launched into an elliptical orbit in the ecliptic plane and sampled the solar wind between 0.3 and 1 astronomical unit. They found that within these distances from the Sun, the density of the solar wind decreased roughly with the squared distance to the Sun as expected, but the temperature of the solar wind decreased much slower than expected. This meant that the solar wind must be continuously heated in interplanetary space. But, on the other hand, the solar wind kept a rather constant speed. Therefore, this heating did not result in a significant solar wind acceleration; the solar wind has to be accelerated much closer to the Sun.

The next solar wind mission, the Voyager satellites, targeted the outer solar system in order to find the boundary of the heliosphere. The boundary of the heliosphere is defined as the surface where the pressure of the outflowing solar wind is equal to the pressure of the interstellar medium, and at this surface, a termination shock should form where the solar wind decelerates to subsonic speeds. In 1977, the twin Voyager satellites were launched on a journey to Jupiter, Saturn, Uranus, and Neptune, and used swing-bys at these planets to further accelerate and to begin their subsequent interstellar mission. It took until 2004, 27 years after its launch, that Voyager 1 crossed this termination shock of the heliosphere at a distance of 94 AU from the Sun, *i.e.*, at a distance which is three times farther out than our outermost planet, Neptune. Three years later, Voyager 2 also crossed the termination shock at 84 AU from the Sun. The next aim was to probe the interstellar medium. They also achieved this goal: Voyager 1 at 122 AU in 2012 and Voyager 2 at 120 AU in 2018. With that, the Voyager satellites were the first human-made objects that left the heliosphere and entered the interstellar medium.

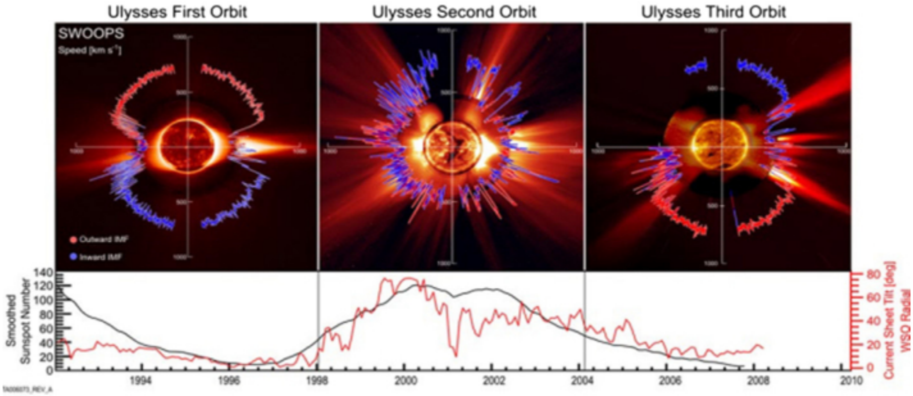


FIG. 4.8 – The three orbits of Ulysses around the Sun, from pole to pole. The curve around the Sun represents the solar wind. The further away from the Sun, the faster it is. The blue or red colour of this wind represents the direction of the magnetic field in the solar wind. The inversion of this field from one cycle to the next between the first and third Ulysses' orbits is clearly visible (credit ESA – NASA – SwRI).

As the previous measurements have been mostly made in the ecliptic, the natural next step was to sample the solar wind out of the ecliptic. To that end, NASA and ESA developed the Ulysses mission. Ulysses launched in 1990, used a gravity assist manoeuvre around Jupiter to be catapulted from its orbit in the ecliptic to an orbit over the poles of the Sun. Then, Ulysses needed about six years, *i.e.*, half a solar cycle, to complete one entire orbit sampling the solar wind over all heliospheric latitudes. During solar minimum, Ulysses found two well-separated magnetic solar wind hemispheres. Both hemispheres were dominated by high-speed streams that originated from the polar coronal holes. The slow solar wind was measured merely in a band close to the ecliptic. During solar maximum, however, the picture becomes much more complex. The slow solar wind and high-speed solar wind streams appeared at both hemispheres with positive and negative polarities; high-speed streams regularly appeared in the ecliptic following their coronal hole counterparts while the slow solar wind was not confined anymore to low latitudes but also appeared at higher latitudes. Therefore, the solar wind also shows a clear solar cycle dependence.

All these missions enabled us to study the propagation of the solar wind in amazing detail. However, there is one important regime for the solar wind which we cannot sample and which is still scientifically puzzling: the solar corona. Nowadays, scientists still debate how the solar wind is accelerated and what the sources of the slow solar wind are. But, due to the immense temperature of the corona of about 1 million K, we cannot simply send a satellite there to take *in-situ* measurements; more advanced techniques must be used. To this end, two further satellite missions have been developed: Parker Solar Probe, in honour of Eugene Parker, who predicted the solar wind, and Solar Orbiter. Parker Solar Probe was launched in 2018 and aims to make *in-situ* measurements as close as possible to the Sun. To withstand the extreme temperatures, strong heat shields were developed. Furthermore, the Parker Solar Probe was inserted into a highly elliptical orbit reaching from about 0.7 AU to a very close perihelion with an extremely fast flyby of the Sun that merely takes a couple of days. Parker Solar Probe completes one flyby every few months, and by using Venus gravity assist manoeuvres, it adjusts its course to get closer and closer to the Sun. In 2021, Parker Solar Probe got close enough to the Sun to dip and take *in-situ* measurements in the magnetised corona for the first time, at a mere distance of 19 solar radii from the solar surface. Solar Orbiter was launched in 2020 and combines *in-situ* solar wind measurements with remote observations of the solar corona. Thereby, the Solar Orbiter will fly as close as 0.3 AU to the Sun to achieve a higher spatial resolution with the remote sensing instruments and to ease the linking between the remote and *in-situ* observations. Solar Orbiter will also shift its orbit out of the ecliptic by up to 22 degrees to get a glance at the Sun's poles. The focus of Solar Orbiter is on the chemical composition of the young solar wind near the solar corona. The solar source region of the slow solar wind has to have the same chemical composition as the *in-situ* measured slow solar wind plasma. Therefore, if we can create maps of the chemical composition of the solar corona, we should be able to finally identify the source region of the slow solar wind. Both missions are still ongoing, but scientists have great hope to finally understand the source of the slow solar wind and the solar wind acceleration process.

Coronal Mass Ejections

Both the Skylab space station and the SMM mission were able to observe coronal mass ejections by using coronagraphs, an instrument already invented in 1931. By occulting the bright light from the solar surface, one is able to detect large-scale structures travelling in the solar corona that appear much fainter than the solar surface. Both Skylab and SMM observed the Sun during a declining phase of the solar cycle, a time when coronal mass ejections are scarcer. The launch of SOHO and later SDO significantly improved our data on coronal mass ejections.

The SOHO spacecraft carries three LASCO (Large Angle and Spectrometric Coronagraph) instruments: C1, C2, and C3. LASCO-C1 was supposed to observe the corona at distances from the centre of the Sun between 1.1 and 3 solar radii, but was lost during SOHO's freezing and communication loss in 1998; LASCO-C2 observes between 1.5 and 6 solar radii; and LASCO-C3 between 3.7 and 30 solar radii. The observations from these instruments are used in daily space weather operations. Similar to EIT, the LASCO images are artificially coloured: not according to the temperature it observes, but according to the coronagraph that is used: C2 in orange, C3 in blue. Later on, for the coronagraph images of the STEREO-A and -B spacecrafts, a similar colour style was adopted.

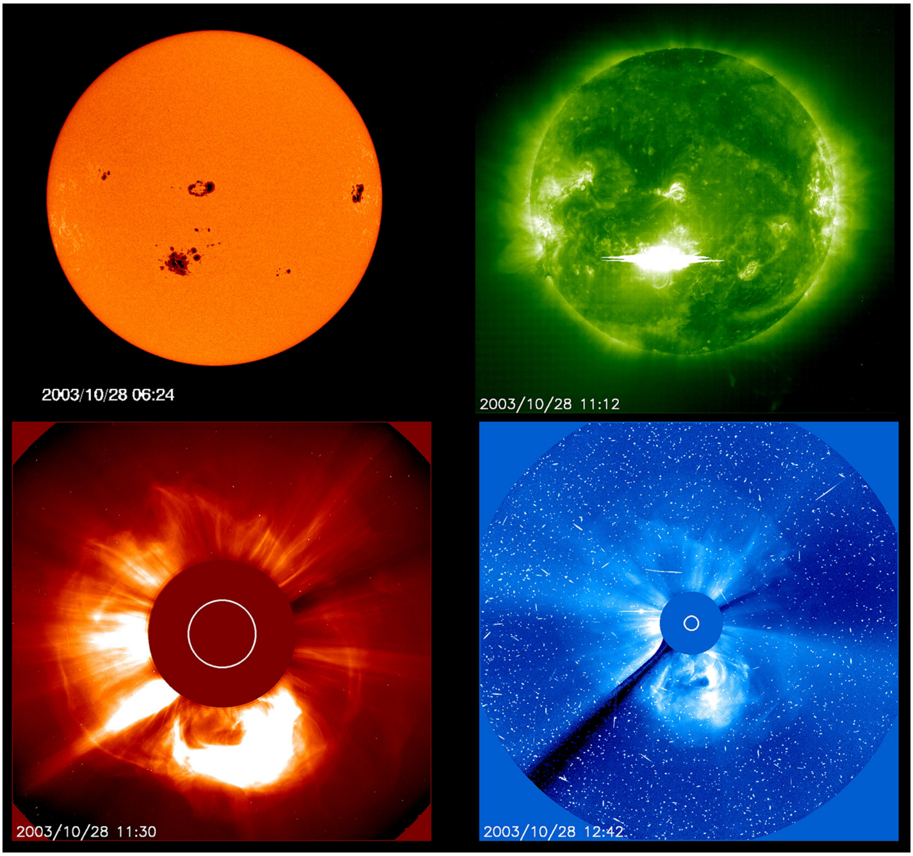


Fig. 4.9 – The solar flare and associated coronal mass ejection of 28 October 2003 seen by SOHO. Left, top: the Sun in visible light observed by the MDI instrument. Right, top: EIT image from SOHO. Bottom left: LASCO-C2 image. Bottom right: LASCO-C3 image.

Combining data from all the different instruments, ranging from EIT to coronagraphs, we are able to track phenomena all the way from the Sun through the solar corona. Coronal mass ejections are often related to solar flares and/or a prominent eruption. The eruptions typically eject a few billion tonnes of material at high speeds of several hundred kilometres per second. If a coronal mass ejection is directed towards the Earth, it reaches Earth within three days on average. However, the fastest coronal mass ejections, with typical speeds of over 2,000 km/s, can take less than one day. As we have seen earlier, coronal mass ejections are linked to the most intense geomagnetic storms observed on Earth.

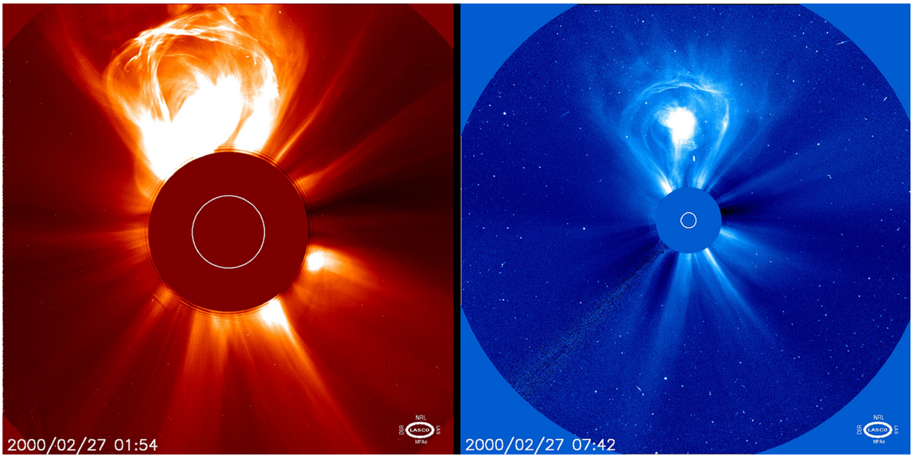


FIG. 4.10 – A textbook coronal mass ejection as seen by LASCO-C2 and LASCO-C3 coronagraphs on board of SOHO with a 6-h time difference. A typical coronal mass ejection consists of three main parts: a bright white front, a dark cavity and a bright core, which is a result of the environment right before the eruption.

SOHO has observed more than 20,000 coronal mass ejections, averaging several coronal mass ejections per day. However, the occurrence of coronal mass ejections is highly dependent on the solar activity level. During solar minima, they are less frequent: about 0.5 coronal mass ejections per day on average. Their speed is also significantly lower. However, during solar maxima, there are about five coronal mass ejections per day on average, about 10 times as much as during solar minima. As a bonus, SOHO has discovered more than 4,000 comets, making it the largest comet discoverer of all time! Some of these comets get too close to the Sun and are vaporised…

The images from SOHO's coronagraphs are spectacular, but in reality, they only show us one viewpoint of a structure that spans widely in three-dimensional (3D) space, nor can we determine whether the coronal mass ejection is propagating toward Earth or away unless the solar source has been identified. As SOHO is located on our side of the Sun, we are not able to observe the far side of the Sun. Knowing the occurrence of active regions on the side of the Sun that is rotating towards us has become important for the forecasting of space weather.

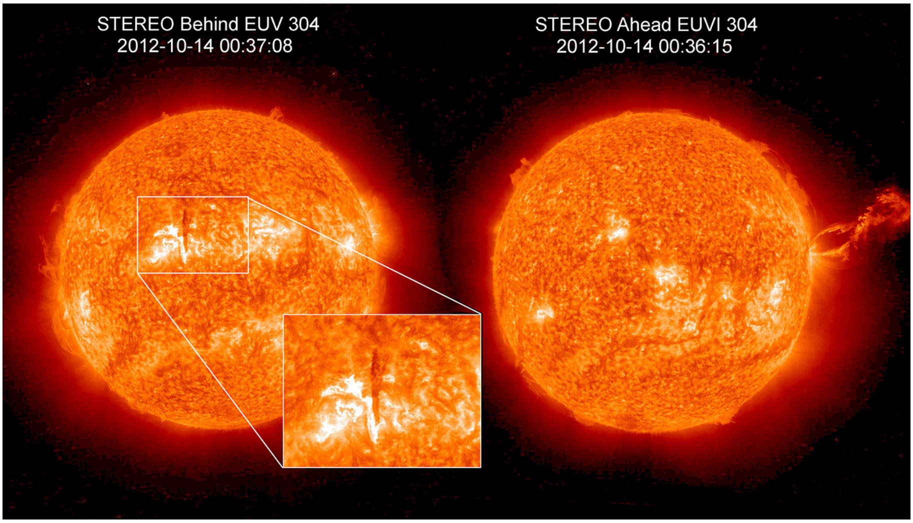


Fig. 4.11 – The prominence observed by STEREO-A (right) and STEREO-B (left). STEREO-B observes the prominence as a dark line in front of the solar disc, the so-called filament; STEREO-A can see the prominence from the side.

To learn more about the 3D structure of coronal mass ejections, and to be able to observe the far side of the solar surface, we need to observe the Sun from at least two different points at the same time, and so NASA's Solar Terrestrial Relations Observatory (STEREO) mission was born. Two almost identical spacecraft, STEREO-A and STEREO-B, were launched in 2006 and placed in two different heliocentric orbits: one that was pulled farther ahead of Earth, and one that was getting gradually behind the Earth, both staying in Earth's orbit but drifting ≈22 degrees per year. Currently, STEREO-A is still operational, but we lost contact with STEREO-B in 2014 right around the time that both spacecraft started the passage behind the Sun.

As the spacecraft had a larger and larger separation angle from Earth, they provided scientists with remote sensing observations of the solar disk and solar phenomena in the corona from multiple viewpoints for the first time. Providing scientists and space weather forecasters with three-viewpoint observations of coronal mass ejections from 2006 to 2014 (STEREO-A, STEREO-B, and LASCO) and from then onwards two-viewpoint observations (STEREO-A and LASCO), that allowed them to learn more about the 3D morphology of coronal mass ejections. With the newly launched PSP, they are able to observe coronal mass ejections in a radially aligned manner (with PSP residing close to the Sun-Earth line) but much closer to the Sun. In contrast, the observations from Solar Orbiter will provide data from an inclination angle of about 25 degrees, allowing for new information out of the ecliptic plane.

Solar Flares

Solar flares have been observed since the Carrington event, but it took a long time to understand them. Flares are intense but localised eruptions in active regions. They emit electromagnetic waves in the entire spectrum to interplanetary space, from X-rays to the radio regime. Flares are classified by their GOES soft X-ray flux in the categories A, B, C, M, and X, where the scale is logarithmic and X-class flares are the strongest category.

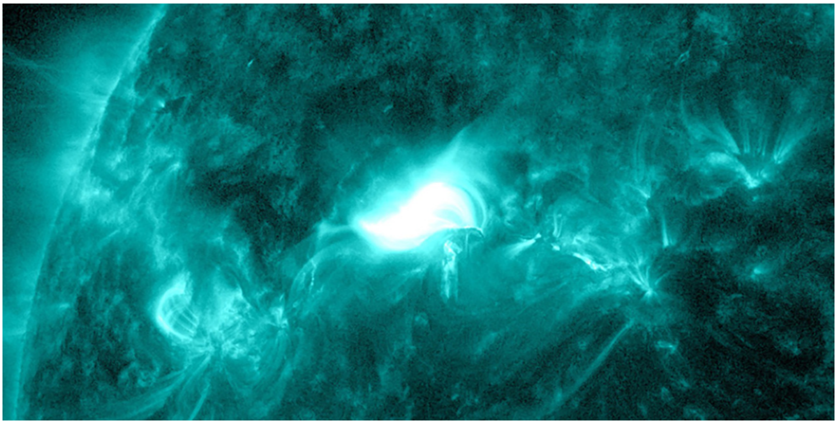


FIG. 4.12 – Image of a M-class flare in AIA-131 on May 3rd, 2023 (credit NASA).

The appearance of a flare is associated with the magnetic reconnection of magnetic loops in the active region. By motion of the loop footpoints in the photosphere and by nearby magnetic flux emergence, the magnetic configuration of the active region becomes more complex, and electric currents form that store free energy. At some point, the active region will release this free energy by magnetic reconnection of nearby loops. At the reconnection site, electrons are accelerated to a small fraction of the speed of light, propagating upwards and downwards along the loops. The electrons propagating upwards to the less-dense corona stimulate plasma emission, which can be observed as Type III radio bursts, *i.e.*, a fast drift from high to low radio frequencies. The electrons propagating downwards abruptly stop when they hit the chromosphere and emit hard X-ray bremsstrahlung. The energy release in the chromosphere results in an evaporation of the chromospheric plasma into the corona, forming post-eruption arcades that glow in the extreme ultraviolet. The X-ray and extreme ultraviolet radiation of flares is absorbed by Earth's atmosphere and does not reach Earth's surface. However, it ionises the upper atmospheric layers, and, by that, changes the height of the ionospheric layers and can cause radio blackouts.

The first flare ever recorded was the Carrington event in 1859. This event was so strong that Carrington could observe brightening flare loops using a simple telescope in white light. White light flares are a very rare phenomenon. With the invention of the spectrograph and the discovery of the chromosphere, brightening flare ribbons were also regularly observed in sunspots in the chromospheric Hα spectral line. As solar flares are related to sunspots, their occurrence rate naturally follows the solar cycle.

With the discovery in the 1940s that the solar corona is a million degrees hot, the Sun became an interesting target for a multitude of instruments. In 1950, Wild and McCready built a radio spectrograph in the metric wavelength regime and pointed it at the Sun. Their recorded dynamic spectra showed several types of solar radio bursts, and their classification of these radio bursts is still used today. In 1960, the first X-ray observatory to observe the Sun was launched aboard the Solar Radiation 1 (SOLRAD 1) satellite. SOLRAD 1 could only measure the integrated X-ray flux over the solar disk but proved that radio blackouts on Earth are related to an enhanced solar X-ray flux that occurred simultaneously with flares. Since then, the integrated solar X-ray flux has been observed by many satellites. The probably most important series of X-ray flux measurements are provided by the Geostationary Operational Environmental Satellites (GOES) series, on which the GOES soft X-ray flare classification is based.

As the solar X-ray emission recorded by SOLRAD 1 was believed to originate in the solar corona, the hunt for solar flares in coronal images started. For that purpose, Skylab was equipped with the Apollo Telescope Mount, and it was able to catch six major solar flares. Since then, the solar corona has been scanned regularly for solar flares, first by the SMM since 1980, then by SOHO since 1995, the Reuven Ramaty High Energy Solar Spectroscopic Imager (RHESSI) since 2002, by SDO since 2010, and by Solar Orbiter since 2020. Each time, the imaging techniques became more refined and the spatial and temporal resolution of the images better, allowing us to observe more and smaller flares and more in the flaring region. Nowadays, far more than 100,000 flares have been observed by all these imagers.

Solar Energetic Particles

There is a further, comparably new scientific subject on space weather and space climate which only started in the 1940s: solar energetic particles (SEP). Solar energetic particles are electrons, protons, and heavy ions that are accelerated up to Gigaelectronvolt energies into interplanetary space. They are accelerated in shocks associated with magnetic reconnection in solar flares and with coronal mass ejections. When solar energetic particles leave the Sun, they roughly follow the interplanetary magnetic field, *i.e.*, the Parker spiral. Wide-spread events in heliospheric longitudes have been reported as well. Due to SEPs' high speeds, they often need only minutes to reach Earth. When SEPs hit satellites, they degrade the satellites' solar panels and can cause damage to the satellite electronics. In Earth's atmosphere, SEPs strike atomic nuclei of the air, producing particle showers of secondary particles similar to those produced by cosmic rays. Rarely, as they have considerably lower energies, the showers initiated by the SEPs can reach the Earth's surface and be recorded as ground-level enhancements by multiple neutron monitors simultaneously.

The first SEP event was recorded indirectly by Scott Forbush in 1942 as a ground-level enhancement in a neutron monitor data. Since then, a network of neutron monitors on Earth as well as energetic particle detectors aboard satellites monitor SEP events. Satellites equipped with energetic particle detectors include the Helios mission, SOHO, Wind, ACE, STEREO, Solar Orbiter, and Parker Solar Probe.

Advances in observations, both from Earth's surface as well as from space, taught mankind more than they could imagine. Our life on Earth is not only influenced by the light from the Sun that we can see with our eyes; there is so much more to unfold. We now know that the Sun emits light in a much larger range of wavelengths than we can see and that our Earth is influenced by solar wind, coronal mass ejections, flares, and solar energetic particles, which all originate from the Sun. Some of the most beautiful phenomena, such as aurora borealis, are a consequence of the interaction of these solar phenomena with the Earth's magnetosphere and atmosphere. With the space age and the rise of satellites, our society has become more and more dependent on technology: electricity, GPS systems, telecommunications, aviation industry, … How are those systems affected by the Sun-Earth relationship?

References

Gosling J. T. (1993) The solar flare myth, *J. Geophys. Res.* **98**(A11), 18937–18949. https://doi.org/10.1029/93JA01896.

https://inis.iaea.org/search/search.aspx?orig_q=RN:8345445.

https://history.nasa.gov/SP-402/ch7.htm.

Chapter 5

The Time of Impacts

That's it! The actors are in place. It only took a few decades for our entire conception of what surrounds the Earth to be swept away to become more subtle, more complex, and more beautiful too. Let us remember that before the twentieth century, we knew only a few forces in the universe, which we thought were inhabited by a continuous ether. We thought that solar energy could only be chemical and that the Sun itself had a lifespan of a few million years. But soon, scientists discovered the foundations of magnetism, on Earth as well as on the Sun. They made the link with the polar auroras. Then the invention of radar revealed our ionised environment and the progress of physics in understanding the nuclear character of our star. As soon as the space age began, the solar wind, theoretically presumed, was confirmed. In the middle of the 20th century, the International Geophysical Year allowed us to understand that the aurora is the most spectacular effect of this wind. But many others were revealed in a few years; the magnetospheric currents, the existence of a hot atmosphere although very empty above 75 km of altitude. Soon, everything is precipitated: SOHO, this space observatory of the Sun, opened the curtain on a dynamic, active star, with changing moods, which vary not only the solar wind but also the radiation. However, our conception has changed. CLUSTER, its magnetospheric counterpart, shows us that all these beautiful diagrams of the magnetosphere that have flourished since the sixties need to be revised because nothing in our magnetic environment is stable.

So astrophysicists dive into the archives. They discover that on the evening of August 28, 1859, auroras were visible as far south as the Caribbean. This strong event cleared the way in interplanetary space for what was coming next. Lord Carrington, who regularly observed the Sun, noted on September 1 a very violent flash of light that began at 11:18 a.m. and lasted several minutes. This flash seemed to have occurred in a place where he had already noted an impressive number of sunspots. Shortly thereafter, power lines heated up, and telecommunication operators saw their Morse code transmission machines spontaneously catch fire. Carrington documented all of this and was the first to make the connection between the blackouts on Earth and the auroras in the lower latitudes.

DOI: 10.1051/978-2-7598-3610-9.c005

And then, much later, during the night of March 13, 1989, a magnetic storm due to a strong solar flare occurred while, in Quebec, the electrical transformers were already fully loaded at the end of winter. The network collapsed, and for nearly nine hours, five million people were without electricity. The world contemplates, dazed, the first great disturbance due to space weather: a discipline is born.

Description of the Impacts

Many of the impacts come from the fact that we depend on satellites on a daily basis. Sending an email can go through one or more of them. Guiding ourselves by GPS, deciding what to wear based on the weather, watching TV from any country from the couch, phoning, texting, surfing the internet, and listening to the radio. But also monitoring the rise of ocean waters, global warming, military troop movements, atomic or volcanic explosions, giving alerts … Or even undertaking technological tests for industry.

Satellite orbits are classified into several types. Geostationary satellites fly at an altitude of just under 36,000 km. There, the centrifugal force that tends to expel them from the Earth compensates for the gravity that brings them back. The geostationary orbit (GEO) has limited applications because it is constrained to the plane of the equator, but it is very popular because the satellites always fly over the same point on the Earth. They are therefore permanent relays for all our communications, our television broadcasts, and classical meteorology. This is the reason why television antennas all point towards the equator: that is where the signal comes from.

Thirty-six thousand kilometres? When the Sun is quiet, the magnetopause – the boundary between the zone in which we are under the influence of the Earth's magnetic field and the zone where we are exposed to the solar wind – is about 60,000 km from the Earth. The geostationary orbit takes place entirely in the magnetosphere, above the Van Allen belts (see chapter 3). The satellites are sheltered from the solar wind and this is one of the reasons why the idea that the magnetosphere "protects" us is often found. But if the solar activity becomes strong, the magnetopause, on the day side, is pushed by the solar wind in the direction of the Earth and descends several thousand kilometres towards the Earth. The satellites are then in the magnetosphere where electric charges accumulate, or are even directly exposed to the solar wind particles when they are most aggressive. The effects can be significant: degradation of solar panels, and parasitic electric currents that disrupt onboard computers. One of the most spectacular impacts is, of course, on telecommunications.

The zoo of orbits is very populated and meets specific needs: do we want to fly over the same geographical area several times a day, do we want to see the poles, do we want to have a global or local view of the Earth… Often, commercial reasons come into play. Thus, meteorology, communications, positioning… each application has its own orbit. The lower a satellite flies, the denser the surrounding atmosphere is and the greater the friction it experiences. At high latitudes, a geomagnetic storm has the consequence of heating the upper atmosphere. Similarly, if a solar flare occurs, the ultraviolet radiation increases, creating a strong excitation of the upper atmosphere on the sunlit side. Heated-up gas of course cannot expand downwards: so the upper atmosphere expands upwards so the satellites see the friction increase. They are slowed down, and dive. To get them back to their flight altitude, they have to perform maneuvers that consume fuel and therefore shorten their functional lifetime. Imagine what this can mean for a fleet of Low Earth Orbit (LEO) communication satellites!

But other damages can be instant and irreversible. The Americans, always keen on sensationalism, have called the fastest particles in the solar wind "killer electrons" – which are often protons, not electrons. They collide with the satellites' hulls, creating an electric spray that spreads through them. If this current reaches the onboard computer, it can completely disable the entire satellite. Most of the time, this is as simple as a reset, just like you would do on your own personal computers. But sometimes, the satellite's attitude control is affected: nothing indicates where the Earth is anymore, and the satellite starts spinning on itself, unable to position itself. A satellite worth several million euros can be lost for a single solar particle!

Other damages, perhaps less violent, can occur. Solar panels, obviously directed towards our star, are the first to be affected by increases in extreme ultraviolet radiation. Remember that the characteristic of this radiation is its ability to be so energetic that it can tear electrons from any atom or molecule, irreversibly degrading them. Whether we use metals, plastics or glasses, nothing can be done about it: the deterioration is inexorable. The situation is even worse for space telescopes observing the Sun!

Solar wind and radiation "conspire" to degrade the electronics of satellites. For those flying below about a thousand kilometres, there is the added effect of atomic oxygen in the Earth's atmosphere. Here again, solar activity has an influence. The more there is, the more this oxygen is ionised. Once ionised, it travels faster, carried by the magnetic and electric fields, and its chemical properties become more efficient in corroding the panels.

These satellites age quickly, and the closer they are to the Earth, the more friction against the atmosphere, which depending on the solar activity, accelerates their senescence. What happens to them at the end of their life? In a geostationary orbit, they are pushed outwards to be expelled into a graveyard orbit? – only a couple of hundred kilometres above. This is a temporary solution: sooner or later they will have to be forced to climb to 600 km above the geostationary orbit so that they can drift slowly before disappearing – in the very long term – into deep space. Satellites in low orbit plunge into the atmosphere where they burn up. The size of a satellite is about the same as a car. The pollution generated by a single satellite is therefore not considerable. However, between 1957 – the launch of Sputnik – and February 2020, about 10,000 satellites were launched. But just after three more years, due to the expansion of satellite constellations, the amount doubled, also making the number of active satellites close to 8,000.

Of the total number of active satellites, the vast majority fly in Low Earth Orbit (2,070 satellites), and the rest are in geostationary orbit.

More than 7,000 satellites have disappeared or are still in space, not working. Evaluating the number of those affected by a failure related to space weather is quite a feat. A few years ago, a study was conducted for the European Space Agency. In the course of the investigation, none of the space safety managers of the space operators would ever admit to a failure. However, in 2008, Mak Tafazoli, from the Canadian Space Agency, tried his own evaluation. According to him, there were 156 failures between 1980 and 2005 on both civilian and military satellites, 25 of which were related to space weather and space climate.

Are the costs disastrous? According to NASA, the average cost of manufacturing and launching a six-ton satellite was about $141 million over the 2012–2020 period. The European Space Agency (ESA) estimates the cost of launching a satellite of the same size at 100 million euros. Nevertheless, 25 satellites were lost over 25 years… One per year… Even at $150 million each, it deserves attention. This is one of the harsh realities of space weather and space climate.

For us, the answer is 'yes', even if those primarily concerned are not convinced. Why 'yes'? The first reason is that as space is increasingly taken into use, the number of satellites in LEO will increase, and at the same time, the use of radiation-hardened components and the size of electronics will be reduced. Although the cost of a satellite will decrease, the vulnerability to space weather and space climate will increase, leaving more useless junk. The second reason is still more general: A space weather event on a Carrington level could, according to several models, bring down not one, but several satellite constellations. Communications, positioning, and a large part of the images would disappear… The exsanguinated space network would generate, on Earth, a catastrophe whose cost, this time, could be ten to one hundred times that of the largest terrestrial cyclones, spread over the whole globe. Satellite customers would turn to space operators, who, in turn, would turn to their insurers. Few insurers currently cover satellites. They can reimburse a few lost units per year, but certainly not thirty or even a hundred. For them, it would be an assured bankruptcy, causing the banks – their main shareholders – to plunge into the red. It is the whole banking system that would be destroyed by a solar flare. We are no longer talking about a few hundred million euros, but thousands of billions. It is therefore not surprising that global insurers, such as AXA XL, have included space weather as an emerging risk. This French insurer, with a global footprint, is now the third-largest space insurer. According to its global technical director, Mr Bousquet: 'since 1980, in the civil geostationary telecommunications satellites on which AXA XL has visibility, 570 have been successfully launched. However, out of this population, 159 satellites – *i.e.* 28% – have had a failure important enough to reduce their capacity in telecommunication service, and 35 satellites are even in total loss and either have been put on a so-called "graveyard" orbit to leave their place to a replacement, or remain in space as a collision hazard for the others'.

Space weather is not, by far, the primary cause. Still, according to Mr. Bousquet, 'the breakdowns of these satellites come mainly from electronic, mechanical or power supply elements. In low orbit 4% are insured while in GEO orbit this percentage rises to 46%'.

These figures are revealing and exciting. In the case of strong solar activity, the magnetosphere on the day side is compressed and moves towards the Earth. Satellites in geostationary orbit then find themselves outside our magnetic cavity, directly exposed to the solar wind. The situation may become even more worrying if we consider, still following Mr. Bousquet, that 'telecommunication satellites carry more and more electronics, to be more flexible and to be able to change missions once in orbit if the market requires it. These electronics are more sensitive to space weather than older generations using analogue components. In addition, the largest satellites use electric propulsion rather than chemical propulsion; this type of system is potentially more sensitive to radiation. Finally, the development of "new space" around many small satellites with less protection than large satellites, less shielding, and less redundancy makes them much more vulnerable'.

The solutions are beginning to be known: duplicating the instruments on board or the satellites, protecting the cabins, saving energy and data. All this has a weight, and each additional kilogram adds thousands of euros. How can we convince operators to spend so much money on a disaster that may never happen?

Another economic argument weighs in the balance of space. It starts with a delicate question: how much is a human worth? To the average person, a human is worth a human, and all humans are equal. Not to an insurer. A footballer's leg, or a soprano's vocal cords, are worth more than ordinary legs and vocal cords. But who is worth more than all the others? A spaceman. The most violent solar flares are lethal to a human in space. It is difficult to protect them behind lead walls. At best, we can tell them to put on their spacesuits and stand behind as many walls as possible. Not very reassuring, especially in the perspective of long stays in space. For example, to go to Mars…

Why are spacemen worth so much? Because if a crew dies, a whole program stops, involving hundreds of thousands of employees in thousands of companies spread over an entire territory.

We are not done with space, and we must now address one of the most impressive aspects. For decades, it has been accepted that a satellite at the end of its life would deorbit and burn up in the atmosphere. On average, one satellite or launch vehicle stage per week enters the atmosphere. This represents only about one-third of the initial mass! Since the beginning of the space era in 1957, there have been about 5,600 launches generating more than 220 satellites fragmented in orbit, or about 8,800 tons of space objects. The 220 fragmented satellites generated more than 34,000 fragments of more than ten centimetres, circulating at about eight kilometres per second, ten times bigger and ten times faster than a bullet from a large caliber rifle, enough to explode a solar panel, to tear off an antenna, or to perforate a spacecraft cabin; no armor can resist particles of more than two centimetres launched at this speed. We have been able to catalog and track 22,300 (as of February 2020) of these objects, 'only', one would be tempted to say, but what about the 900,000 fragments of one to ten centimetres, and the 128 million smaller objects? They circulate haphazardly like so much dangerous dust above our heads.

Space is gigantic, and the probability of an encounter remains low. However, it grows exponentially with time. Thus, it is estimated that there have already been five hundred failures, collisions, or abnormal events resulting in fragmentations and debris. Recently – in January 2007 – the number of space debris increased by 25% when China destroyed one of its own meteorological satellites, FengYun-1C, using one of its own space projectiles. Only the USA had carried out such a voluntary space destruction, in 1985, attacking one of its solar wind observation satellites. Both nations are flexing their muscles at the risk of filling our environment with debris, of which they themselves could be the victims.

How is this related to space weather?

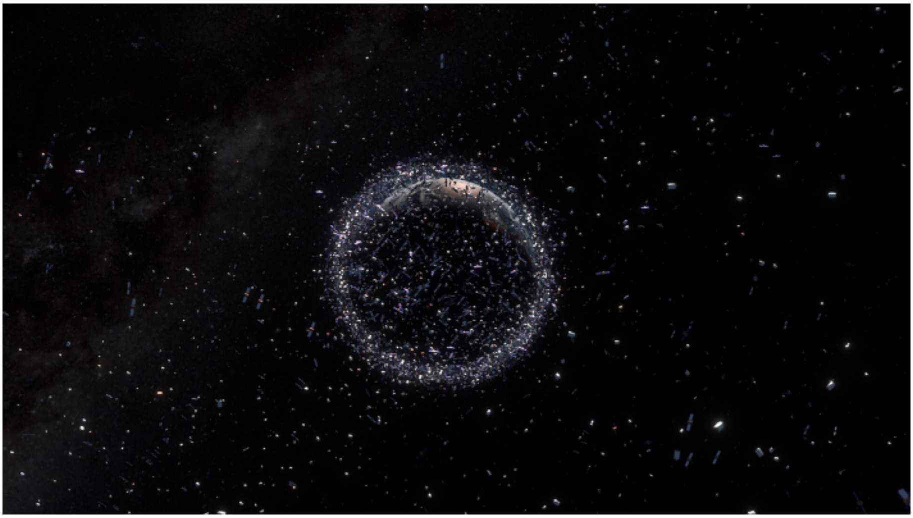


Fig. 5.1 – Representation of satellites and space debris in low orbit (credit CNES).

When a solar flare occurs, the strong emission of X-ray and extreme ultraviolet radiation that accompanies it heats the upper atmosphere on the dayside. The atmosphere expands and influences the orbits of the debris within it. The two global tracking centres – NASA's Langley and ESA's Darmstadt – lose sight of them. Their trajectory is now distorted by the currents of the thermosphere, with the risk of damaging other satellites or, worse, the International Space Station. The movie 'Gravity' has widely popularized this theme. It is serious and real. So, what can be done? It took years for the UN to address this problem. It alone had the legal authority to do so. It created the Committee for the Peaceful Use of Outer Space (COPUOS), which drew four lines of force: to know the situation, to protect oneself, not to create debris, and finally to clean up space. The first two points are undoubtedly the prerogative of space meteorology! Protecting oneself requires, for example, being able to calculate the orbits in case of a random re-entry during a solar event. For this, it is better to be warned in advance that a flare is going to occur, because the radiation, once emitted, takes only eight minutes to begin its deleterious effects.

Fig. 5.2 – Space debris photographed from the space shuttle (credit: NASA).

Even if it is no longer the domain of space weather and space climate, ESA now follows the Zero Debris approach on its spacecraft and suggests that its partners follow such a charter. Additionally, the cleanup of space debris raises an important legal point. Suppose China retrieves a decommissioned Russian satellite, stuffed with electronics and, who knows, classified data. Who owns it? According to the laws of the sea, it belongs to China. But does that apply to space?

In 2008, France passed its own space law that includes binding requirements on space debris. After a transitional phase, this law has been in force since December 10, 2020. A satellite can only be launched if, during its complete life and its decommissioning (not launched), the probability of causing at least one casualty is less than or equal to 0.00002. Is this low? That is one death for every 50,000 inhabitants, a little more than a thousand in populated areas like most of the western European countries. In the case of an uncontrolled re-entry, this number is multiplied by five. So, it is helpful that space meteorologists are improving their predictions!

One of the most common applications of satellites is positioning. Your car now shows where you are and guides you to your destination. As we saw in chapter 3, it needs to be connected to at least four global positioning satellites (GNSS in jargon) to do this, and space weather forecasters use the discrepancies between the calculated position on the ground and the actual position to deduce the total electron content along the wave path.

But if a solar or geomagnetic disturbance occurs, the electron content of the ionosphere changes. In turn, this changes the opacity of the air for the waves. Instead of coming to you in a straight line, they are deflected, and your estimated position is in error. You have all experienced it: suddenly, the screen in your car positions you next to the highway you are driving on. You laugh: ‘this GPS is crazy’. No, it’s not; an electronic cloud due to a burst of solar activity has just passed between you and one (or more) of the positioning satellites. So you know it’s a horizontal error of a few metres. This is not a big deal, you think, because you know perfectly well that you are on the highway. But if your car were driven by a robot guided by this positioning system, how would you feel? Or if you were one of those truck drivers in the foggy Siberian tundra?

More sophisticated navigation systems use countermeasures to avoid these errors. As the ionospheric signal propagation delay depends on that signal frequency, the navigation satellites transmit multiple different frequencies so that the receivers can make corrections that give positioning accuracy of tens of centimetres horizontally, but still several metres vertically, up to eight in the case of a big magnetic storm. Again, this might seem ridiculous. It is not. A few metres vertically is the difference for a missile between a target and the school next door… Even with multiple frequencies, some applications are still subject to the vagaries of space weather and space climate. Grenoble, France, is located on a plain at the foot of a mountain range. Numerous hydroelectric dams dominate it. If one of them breaks, sirens will sound in the city, and the inhabitants will have to evacuate as soon as possible. These dams are therefore constantly monitored, thanks to GPS-type positioning sensors. It goes without saying that it is better to know the meteorology of space than to evacuate 400,000 people because of a solar flare that could mislead the measurements…

FIG. 5.3 – Artist’s view of four positioning satellites of the European Galileo constellation (credit ESA).

The Sun, as we have explained in previous chapters, constantly emits energy in the form of electromagnetic radiation (most variable in the ultraviolet, extreme ultraviolet, and X-ray bands) and charged particles into space. The latter produces currents in the magnetosphere and electric fields, which in turn, disturb the external magnetic field. The French scientist Laplace (1749–1827) already claimed to be able to detect an aurora borealis from Paris with a fairly accurate compass. But at that time, neither electricity nor radiation was mastered. Today, we use them daily. It is not surprising that solar activity impacts us more and more, sometimes in an indirect and surprising way. For example, when the Sun is active, it produces more solar wind that spreads cosmic radiation. Conversely, we are more exposed to this cosmic radiation, up to very low altitudes, when the Sun is quiet.

The high-energy electrons or atomic nuclei of cosmic radiation are mostly stopped in the atmosphere above 10 km of altitude by collisions with the air, as noted in the third chapter. But these collisions are so violent that they produce a spray of so-called secondary particles (neutrons, mesons, and electrons). At sea level, their quantity is reduced by a factor of 300, but only by a factor of three at an altitude of 9 km.

Electronics are everywhere and they are becoming more and more miniaturised: they are more and more sensitive to this secondary radiation, being less protected by matter. Trains and cars are full of them. It is difficult to obtain information on breakdowns due to space weather. Some information is sometimes filtered out, most often when investigations force the publication of the results of the studies. Train breakdowns in Germany in the early 2000s occurred when they emerged from a tunnel. This was obviously very detrimental to the traffic. It was a component that was too sensitive to cosmic radiation. It seems that the same problem had been encountered in Russia, even if no official communication was provided.

Let's continue to get closer to the ground. The reader surely remembers that the upper atmosphere is composed of the ionosphere – its ionised part – and the thermosphere – its neutral component. Each electron density has its own frequency, which is called the 'plasma frequency'. As the electron concentration varies with altitude, the plasma frequency also varies naturally. Its frequency range corresponds to High Frequency (HF) emissions in telecommunications, typically around 10 MHz.

What happens if we send a wave of this frequency to the ionosphere? It resonates with the ionosphere, ricochets off it like a mirror, and returns to the ground. That's all it takes to communicate at a distance. A transmitter, a receiver thousands of kilometres away, and … a good knowledge in real time of the ionosphere. This principle is not only of interest to radio amateurs. When flying intercontinental flights across the pole, airlines also rely on HF radio to communicate with the outside world. Even with modern satellite links, HF is needed in very remote, high-latitude locations. With space weather, the polar regions may experience long-duration periods of absorption causing radio blackouts, or simply problems with signal propagation owing to large electron density gradients within or inside the auroral oval. The military also often uses it in the field of operations, especially at sea where the transmitter can aim very low without being hampered by terrain obstacles. On land HF is often used vertically, to communicate from one valley to another. It is not surprising that the American army has a space weather battalion, while in France, Major Lionel Birée was able to create an operational military space weather centre. There is no doubt that all the major armies have set their own space weather facilities, although they hardly communicate about them. The army is obviously also interested in all the space impacts described above, but it has its own needs. For example, it has developed a radar that uses the bounce of a wave on the ionosphere to detect a missile on the other side of the horizon. Needless to say, forecasts of geomagnetic activity must be accurate. One cannot risk retaliation if the so-called missile detected is, in fact, a plasma bubble moving in the ionosphere. This is why the detection of these currents and the local prediction of electron concentration variations are so important.

There are many ground-based impacts of space weather. The one that has been most popularised can be considered as the birth certificate of the discipline… In one and a half minutes, a 9,500 MW transformer wiring literally melted in Quebec in 1989. What happened?

The magnetosphere is, as we understood in chapter 3, a place in which many electric currents circulate. It is true that the electron and ion density is small, but the distances are so great – several tens of thousands of kilometres – that their average intensity is several million amperes, that is, ten to one hundred times greater than the currents in a lightning bolt. These currents also flow into the ionosphere and horizontally along the auroral ovals. From close by, through the electric field they produce, they force the electrons circulating in the conductive soil to follow their movement, creating induced geomagnetic currents of low intensity, but nevertheless dangerous. Long and metallic structures are affected by these induced currents: railroad cables, power transmission lines, pipelines, and so on. At the ends of these are power plants and transformers with grounding where the currents will seep in. When, for example, transformers receive a direct current, or a slowly varying current as in the case of geomagnetic induced currents (GIC), they heat up. They are designed to absorb without damage-induced currents of low intensity or disconnect from the grid. However, they are not able to cope with the consequences of large geomagnetic disturbances. When a transformer gets damaged or disconnects from the grid, the rest of the grid, especially if it is under a large load, will not be able to sustain operations, and a cascade of disconnections may occur. Space weather blows the fuses! It should be noted that such currents can have an intensity of more than thirty amperes for tens of minutes.

In the United States, in 1992, during a magnetic storm, an increase in the temperature of the transformers from 60 °C to 175 °C was noted. But the heating can be so significant that the coils literally melt, as was the case in 1989 in the USA and Canada. This is an important matter: we are not talking about small transformers as you use them for your daily objects, but about among the biggest manufactured objects on the planet: 9,500 MW in the case of the one that was destroyed in 90 seconds in Quebec. The costs are considerable. The blackout of March 13 and 14, 1989, plunged five million people in Quebec into darkness, creating formidable difficulties in hospitals, retirement homes… The cost is estimated at two billion US dollars. A transformer in the USA was put out of service – twelve million dollars – while two others in Great Britain were damaged. Incidentally, 1,600 satellites were momentarily 'lost' in the space environment…

To counter this, we resort to surveillance. Electricity companies in countries with conductive soil – in France, this is only the case in Brittany – have multiplied the number of electrical sensors. When a wave of current breaks and heads towards a power station, it only takes a few minutes to lighten the load it, reduce its production, and … buy power from neighbouring countries if they are better sheltered.

Other effects of space weather may seem surprising. Who would think that oil drilling depends on solar activity? And yet… We all have in mind the drilling towers of the oil fields of the Gulf countries: forests of derricks several tens of metres high. The oil platforms and maritime monsters are also well known. It is normal to keep the drilling equipment in one place and start drilling vertically, but then deviating horizontally into the underground target. Horizontal drilling is done to increase the production of hydrocarbons from the reservoir, to avoid hitting undesired geological structures or other wells. Even when accidents happen, with blowouts creating huge environmental disasters, the relief well drilled to stop the blowout is normally a horizontal well drilled exactly to hit the other. How do you find your way underground? The answer seems obvious: you take a map, point in the right direction, and go straight. Hmmm… Going straight underground is not so easy. Since global positioning does not go underground, the best way to proceed is to use the Earth's magnetic field as a compass. Oil company geologists have developed the most accurate field maps in the world, which they don't even share with scientists. Their problem is that if a geomagnetic disturbance occurs, the field is disturbed by several hundred nanoteslas. This small one percent change can lead to an error on the pointing axis that can ultimately be very costly. Here again, the exact value of the loss is a well-kept secret of the industry. They send their specialists to our congresses, make contacts, and sometimes sign research contracts with a space meteorologist under the seal of confidentiality. This is an important problem for them. While waiting to be able to quantify the modifications of the internal field due to a magnetic storm, they need a fine nowcast of the activity. Will they stop when geomagnetic activity happens? No, it is too expensive to stop; think about the hourly rent of their oil rig! They keep drilling but correct their measurements in real time with data from nearby magnetic observatories or models. The impacts of ignoring space weather can be severe, both from loss of income or inability to prevent disasters, costing millions of dollars.

Oil companies are not dependent on space weather only for drilling. Another impact is the corrosion of their pipelines (oil and gas pipelines). These are made of a lot of metal, generally good conductors of electricity. However, electrons flowing along the metal structures, as a result of induced currents generated by geomagnetic storms, cause corrosion. How can this be avoided? By keeping the electrical potential slightly negative in relation to the ground (−0.85 V). What a waste of energy, especially since the pipelines run for thousands of kilometres. So, of course, a very low potential is maintained, just enough to repel electrons from the ground.

Let a strong solar flare occur, and the induced geophysical currents increase. The electric potential of the ground becomes too high; it reverses in relation to that of the pipelines, which the electrons attack and corrode. It is, therefore, necessary to monitor thousands of kilometres of pipes, especially when the Sun is active. The cost of such monitoring is not negligible: in Alaska, operators have to be picked up from their backyards, dropped off on a section of the pipeline (1288 km) of which they can only travel a small portion during the day, and returned to pick them up in the evening. For this particular pipeline, however, the most important damages are due to American citizens who practice shooting...

Aviation is not immune to solar hazards either. It was a French heliophysicist, Pierre Lantos (1942–2007) who, as early as the 1980s, had the intuition that flight personnel could receive significant doses of charged particles, especially when passing under the auroral oval. Let's remember that when the Sun is quiet, it also lets through more cosmic radiation, which creates secondary particle sprays that are still numerous at long-haul flight altitudes. Lantos set up a project he called Sievert, named after the unit that evaluates the biological impact of human exposure to ionising radiation.

In nature, the average radiation is 2.4 millisieverts per year. During a radiological examination, we receive about 0.7 millisieverts. The people who worked in Chornobyl immediately after the explosion of the power plant received the lethal dose of 2 sieverts. Spacemen should not receive more than 1.5 sieverts in their entire career. This is what some solar events produce in less than a day. In August 1972, the radiation dose received on the Moon from a solar storm was estimated at 7 sieverts per hour. By an incredible chance, no Apollo mission was flying that day: the astronauts would have all died, having received more than 15 sieverts in a few hours. Pierre Lantos wondered how much of this radiation remained at transatlantic flight altitudes. Here, the international standards are even more drastic: one should not receive more than 100 millisieverts over five consecutive years, and no more than 50 in one year. His first step, in agreement with Air France, was to equip flight crews with small sensors attached to their uniforms. It was an approach that did not take psychology into account. It generated anxiety, and these sensors were quickly "misplaced": what would happen if the dose was exceeded? Would the employees be banned from flying? The sensors were placed on board the aircraft and a major modelling effort was made to calculate the theoretical dose to personnel on a large number of flights. It was estimated that in September–October 1989, a dose of 100 microSieverts per hour was received at Concorde flight altitudes, and 20 microSieverts per hour at more conventional altitudes. By modelling, it is estimated that in 1956, a solar event could have irradiated flights with doses even 1000 times higher. We remain below the limit in general, but this concern is important enough to constitute permanent safety control.

Perhaps more worrisome is the dependence of air networks on space weather.

To monitor airspace, airports use two types of radar. The "primary" ones are used to locate aircraft in an airspace. Secondary Surveillance Radars (SSR) send coded requests to aircraft and retrieve auxiliary information in return: identification, external pressure, and for some systems, selected technical parameters. If the aircraft does not have the equipment to communicate with them, the secondary radars simply do not see them. To request information from an aircraft, they use the 1030 MHz frequency. To receive it, they use 1,090 MHz.

On November 4, 2015, the Sun experienced a surge of activity that manifested itself in a series of flares accompanied by strong radio wave emissions. Coincidentally, they were concentrated around 1 GHz, precisely the band used by SSRs. Space weather is still unable to predict these radio bursts, let alone the frequency range in which they will occur. When they are detected, it means that they have arrived on

Earth. All secondary radars that were pointing at the Sun at that time were affected. This affected Western Europe.

In Belgium, ghost aircraft echoes occurred between 3 and 4 pm (local). The computer system correctly identified these false alarms and filters them out. Incidents were also reported in Greenland, disrupting the landing of an aircraft. But it was in Sweden that the effect was the most striking. The secondary radars became totally inoperative. Air traffic controllers, deprived of information, were forced to reduce aircraft movements, divert some, and prevent take-offs. Stockholm soon experienced paralysis, creating a traffic jam that spread across northern Europe. Due to a few seconds of solar emission, the European sky was partially paralysed. Airplane delays and passenger diversions cost airlines hundreds of thousands of euros. What would be the cost of a longer, more intense event?

Space meteorology is actively looking for ‘precursor’ signs. A radio burst, for example, occurs outside the impulsive phase of a flare and often arrives about thirty minutes after an X-ray emission. It is often associated with a particular structure of the solar magnetic field. But we have only a few years of observation of the Sun, too few to draw a generalisation. It must be recognised that for the moment, we are unable to predict new radio bursts.

Tourism is of course another potential client of space weather. Auroras have become incredibly popular. Tour operators are willing to bring tourists to see the polar lights if they are certain to witness them. Even with a warning, one or two days in advance. Northern countries have understood this, Finland in the lead, which has equipped itself with observation posts well heated in glass bubbles.

Fig. 5.4 – Polar aurora at Skibotn, Norway, February 2023 (credit: Gaël Cessateur/IASB).

We wonder about our responsibility as scientists. In a few decades, we have seen the glaciers retreat, the snow disappear, we have experienced the rise in temperatures, and the beaches of Tromsø (in the North of Norway, beyond the polar circle) resemble those of the French Riviera. Is mass tourism still sustainable? But at the same time, what right do we have to keep the polar lights out of reach? Why forbid the general public from sharing our emotions in front of their fabulous display?

This naturally leads us to the role that solar activity could have on the climate. Solar activity has always been the main argument of climate sceptics. According to them, it has been more intense than before during the decades that have seen the global temperature of the Earth's atmosphere increase. So why worry? It's natural, there are cycles, and it will return to how it was before, so let's continue to pollute with glee… Except that things are not that simple. The amount of energy that varies on the Sun is of the order of a thousandth of the total, barely one watt per square metre at the Earth level. Most of it is absorbed in the ionosphere and the thermosphere. How can we explain that such a small amount of energy propagates over more than a hundred kilometres, and amplifies to create such considerable upheavals? The orders of magnitude are of the same order as claiming that lighting a candle in Brussels could heat the city of Lille in France. One has to take a deep breath to assert such a 'truth'…

First, let us specify that the question at the heart of the debate is that of solar variability. It is not new that the Sun is the main source of energy for our atmosphere. But what about its variations?

Let's remember that most of the solar energy reaches us in the form of electromagnetic radiation with an integrated average of 1361 watts per square metre outside our atmosphere. This amount, the solar irradiance, varies by about 0.1% over the course of a solar cycle, the average period of which is eleven years. The direct climatic effect of such a small modulation is of the same order of magnitude as the natural variability of the climate, which makes it very difficult to detect in climate data. This is why it has long been given the misleading name of solar constant. Today, we prefer the much more accurate name of total solar irradiance (TSI). Its long-term variation is still poorly known and is the subject of much controversy.

However, the energy contribution of the Sun cannot be reduced to this parameter alone. Most of the solar irradiance is made up of visible and infrared light, which is mainly absorbed at ground and ocean levels when it is not directly reflected back towards space. However, as we have seen, solar variability is essentially manifested by variations in ultraviolet radiation, absorbed in the upper atmosphere, which can ionise or excite its constituents. This component represents only 2% of the TSI – a little more than two watts per square metre – but its variability is much higher, which gives it a significant leverage effect. One of the well-documented consequences is the modulation of the ozone concentration in the middle atmosphere, between fifteen and thirty kilometres. The climatic impact of this modulation has long been neglected. However, numerous observations and numerical simulations show that such couplings are possible and could – the conditional is important – offer solar variability a lever on the climate.

In parallel to this radiative input, the solar system is bathed in a solar wind made up mainly of protons and electrons, as well as in the solar magnetic field. If their energy contribution is very low – about 2 watts per square metre at the Earth level – on the other hand, they vary more or less in phase with the irradiance. Every ten to thirteen years, these various quantities pass through a maximum of intensity.

These variations in solar activity have the indirect effect of exposing the Earth's atmosphere to a variable flux of energetic particles, some of which have speeds approaching that of light. Most of them come from our immediate environment, where they are accelerated and trapped in the Earth's magnetic field. The most energetic ones come from the deep universe: this is, once again, cosmic radiation. These particles ionise the air in their path, giving rise to new chemical compounds. On this question, two scenarios are of particular interest to climatologists.

The first involves charged particles from the near-Earth environment. During periods of high solar activity, the Earth's magnetic field is more disturbed. This results in an acceleration of these particles to high energies, but also, and more occasionally, their precipitation in the upper atmosphere (above sixty kilometres). If these events are very intermittent, on the other hand, their rhythm is modulated by the solar cycle. These particles produce nitrogen oxides (NOx) which affect atmospheric chemistry. Through a chain of mechanisms, this ends up affecting the vertical movement of air masses in the polar region, which influences the regional climate. The impact, on average, could be comparable to that of ultraviolet radiation.

The second scenario directly involves cosmic rays. A theory that has had incredible popularity in the world proposes that the modulation of this radiation by the solar cycle affects the cloudiness, that is to say, the rate of cloud cover: when the Sun is calm, it emits less cosmic radiation, which creates nucleation nuclei at low altitude, on which water vapour condenses to form droplets and, therefore, clouds. This theory, proposed in the 1990s by Svensmark and Friis-Christensen, finally provided a mechanism linking solar activity to climate! Since the Sun was active during the last thirty years, there was less cosmic radiation and therefore fewer clouds and the global temperature increased. The international community welcomed this proposal with joy and even a certain benevolence. Its authors were invited from all over the world to present it. Unfortunately, a series of errors were highlighted in their calculations: wrong solar index, wrong sample of the global temperature of the Earth, error in the mathematical statistics. By repeating their work with more proven methods and better-calibrated data, their theory collapsed. However, it was so attractive that CERN developed, in the 2010s, the CLOUD (Cosmics Leaving Outdoor Droplets) experiment for the effect of cosmic rays on the atmosphere using a fog chamber. CLOUD has indeed confirmed the existence of a mechanism of aerosol nucleation in conditions close to those prevailing in the lower atmosphere. However, neither the aerosol microphysics nor the cloud cover observations allow us to attribute a significant climatic impact to this scenario: the cosmics would rather produce cirrus clouds, high altitude clouds, whose effect is very weak, even contrary to that proposed by Svensmark and Friis-Christensen.

As the years went by, the rising trend of solar activity reversed. Since the beginning of the 2000s, solar activity has declined: we should therefore have more clouds and a decrease in temperature. This is not what is observed. The similarity between the evolution of cosmic rays and the evolution of cloud cover was therefore more of a coincidence. Too late! The idea seduced the public, the debate became poisoned.

Being wrong in science often happens. All the natural sciences proceed by making propositions that we spend an extraordinary amount of time refuting. The most solid theories, accepted by the scientific community, are those that resist refutation. Svensmark and Friis-Christensen had a magnificent idea, for which they must be congratulated. It has opened up a field of experiments and increased our understanding of the climate system. In particular, one aspect of this scenario is currently attracting renewed interest. It is the presumed impact of the global electrical circuit on the movement of ions in the lower and middle atmosphere. The Earth behaves like a giant capacitor whose plates are respectively the ground and the ionosphere. The neutral atmosphere located between these plates is traversed by a very weak downward current of the order of one picoAmpere per square metre, which is compensated by the discharges of lightning. This electrical circuit, which feeds the thunderstorm activity, is sensitive to the conditions of the interplanetary environment, which is in turn influenced by the solar activity. We have here an additional mechanism that directly couples solar variability to the lower atmospheric layers. Its influence on the climate remains however minor.

Some indirect mechanisms act in a regional way. For example, a change in ultraviolet radiation due to solar activity seems to have had a particular impact in the past on the circulation of air masses in the North Atlantic region, with a stronger climatic impact in Europe. Here again, this simple statement masks a whole chain of mechanisms whose study is an exciting challenge. The importance of the regional signature of certain global forces constitutes a new paradigm, largely ignored until now, which underlines the risk of reducing the Sun-climate link to global variables. Such regional effects could explain why Europe experienced a cold and wet episode at the end of the 17th century that coincided with a period of low solar activity, called the Maunder Minimum, between 1645 and 1715. But they remain hypothetical. The cooling seemed indeed to have started before the decrease in solar activity. The difficulty of counting the number of spots, which we noted in chapter 2 and which we will discuss again in chapter 6, tends to show that this minimum would not testify to a significantly calmer Sun than the one we had during the first decade of the 21st century. In short, the concomitance of the Little Ice Age and the Maunder Minimum could be due to chance.

Another advance concerns the supposed synchronisation of the solar cycle with natural climate patterns. The Earth's climate varies constantly. Some of these variations, such as the El Niño oscillation, are quasi-oscillatory in nature. Recent studies suggest that natural climate modes (notably the North Atlantic Oscillation) may occasionally synchronise with the solar cycle. The length of available observations does not allow us to say whether this is a coincidence or a physical mechanism, but this synchronisation offers an explanation for the occasional presence of correlations between some climate data and the solar cycle. To confirm this correlation, we must … wait. Wait until we have decades of data, and more active solar cycles than the one we just went through. If it is confirmed, it will challenge the view that a small perturbation of the climate system necessarily induces a small response from it.

But we are not there yet. All these studies depend on observations, the lack of which is a clear problem. Continuous measurements of the solar spectrum, for example, did not really start until 2003, offering insufficient hindsight to draw strong conclusions about variations on decadal and centennial scales. There are therefore two ways to better understand the impact of solar variability on climate. One is based on contemporary observations and relies on a detailed understanding of the physical and chemical processes of the Earth's atmosphere, which is affected in many ways by the Sun. The other takes advantage of natural archives, thanks to which we now have access to several hundred millennia of turbulent solar and climatic history. In this sense, the importance of taking on the national responsibilities of contributing to this international effort of maintaining long-term, distributed observations of local and remote processes in the geospace and beyond cannot be over-exaggerated…

To date, we are able to make a first observation: there is no single mechanism, but a set of coupled mechanisms, whose signatures are frequently similar. Some of them can also compensate for each other. Several have been extensively studied, while others remain to be explored. Only a resolutely multidisciplinary approach, bringing together different scientific communities, can correctly describe this chain of processes and understand the role of each. In the current state of knowledge, several of these mechanisms have a measurable impact on climate, but none of them is a serious candidate to provide a significant contribution to global warming as observed since the 1950s. In comparison, the dominant role of greenhouse gases is indisputable.

How is it then that such doubt persists about the alleged role of the Sun in global warming when the overwhelming majority of scientists first note a dominant effect of greenhouse gases? The solar cycles observed during the latter half of the twentieth century were more intense than in the early twentieth century. The coincidence with global warming is disturbing; climate sceptics see in this correlation one of the proofs of the preponderant role of the Sun in global warming. However, the solar cycle amplitudes in the 21st century have decreased while global warming temperatures have continued to rise.

This works as a textbook example that such coincidences or correlations only have scientific value if they are based on explanations involving credible hypotheses that the scientific community will be able to test, and eventually validate. It is therefore necessary to propose one (or a set of) physical or chemical mechanism(s) that explains how this variation of solar energy can ultimately heat the entire lower atmosphere. Hundreds of researchers are working on this, but the obstacles are numerous. First of all, how do we characterise solar activity with multiple manifestations? None of the proposed observable data is totally satisfactory, and most of the measurements start with the space era, which is ridiculously short compared to geophysical scales. Similarly, how do we characterise climate variation?

There is a more subtle difficulty: the application of statistical methods. The most common method consists of correlating a solar observable with a climate variable. However, correlations do not provide information on causal links and are only meaningful if they are duly validated, notably through statistical tests.

But, perhaps even more important, our scientific community is confronted with sociological behaviours, political statements and economic interests.

It is therefore necessary to open up to the work of others – particularly in the cultural sciences – and to build a common body of definitions. The main difficulty in the study of the Sun-Earth system lies in its enormous complexity, while we lack the information to test the hypotheses. Observations of the climate system have only really started with the space age, which gives us little hindsight to understand what is happening on time scales longer than a decade. A typical example: at the dawn of the 21st century, no one suspected that the Sun would enter a phase of reduced activity. Today, this has become a reality.

However, it is on the basis of patient work of several years gathering dozens of researchers from various backgrounds horizons that we were able to make the diagnosis expressed above, refuting solar activity as a significant actor in the current global warming. The list of mechanisms by which solar variability influences the climate is obviously not fixed; new ones may be discovered tomorrow. On the other hand, in this debate, which is too often polluted by striking statements (often all the more publicised because they are difficult to verify), we wish to affirm the primacy of a scientific approach in which the statements are supported by hypotheses that can be verified and therefore disproved. Any other approach, based on suppositions or just correlations, can only be, from a scientific point of view, an ideological posture.

The Worst Case: Should We Fear Space Weather?

All the above shows that the impacts of a space weather event are numerous. Nevertheless, the sums involved, even if they are in the millions of euros, may seem derisory compared to the global economic flows. Is it necessary to launch into satellite programmes where each element costs a few million? Should we finance forecasting centres, salaries, models, and fast computers? Is it not enough to prepare a few shelters, a few backup procedures and to continue to finance a few researchers to increase knowledge? After all, our technological society has already gone through several solar cycles and, apart from the big blackout in Quebec City in 1989, the effects were relatively small: 2 billion dollars in Canada, 12 million in the USA, a few million for two transformers in Great Britain, and to recover some 1600 satellites. According to the first statistical studies, the Quebec storm has a probability of recurring every one to one hundred and fifty years. The immediate question that economists may ask is: 'Isn't it enough to spend two to three billion in repairs every hundred and fifty years for a solar storm, rather than tens of billions to protect against it?' However, there is another, more relevant question: Won't we be forced, one day, to spend much more?'.

To answer this question, we started to look for the worst case and its consequences on our societies. The first one we found was the famous Carrington event of 1859. We have described it in the previous chapters. Its effects were observed on the aurora and telegraphic telecommunications. What would happen if it occurred today? In 2008, the Japanese Academy of Sciences convened a symposium on this question. The answer is rather chilling: the cumulative damage would reach the pharaonic sum of 2 thousand billion dollars, the GDP (Gross Domestic Product) of Italy! Based on this study and others from the academic world as well as insurance companies, the OECD took up the problem. A report published in 2011 classified geomagnetic storms among the five global risks, on a par with systemic financial risks, cyber risks, social unrest and … pandemics.

Two years later, Lloyd's published a most alarming report. Lloyd's is an international banking and insurance company. Its cost estimation methods are confidential, and likely influenced by certain intentions. The conclusions were as follows: A new extreme geomagnetic storm at the level of the Carrington event is almost inevitable. Historical auroral records suggest a return period of 50 years for Quebec-level storms and 150 years for very extreme storms like the Carrington event. As North America's electrical infrastructure ages and we become increasingly dependent on electricity, the risk of a large-scale blackout increases with each peak in the solar cycle and the danger from geomagnetic storms is one of the strongest. In the United States, it is particularly important along the Atlantic corridor between Washington DC and New York, Midwestern states such as Michigan and Wisconsin, and regions along the Gulf Coast. Twenty to forty million Americans could experience an extended power outage from a Carrington-type storm for sixteen days to two years. The total economic cost of such a scenario is estimated to be between six hundred billion and two thousand six hundred billion US dollars. Weaker storms could result in a small number of transformer outages – about 10–20 – with significant potential damage. The failure of a small number of transformers serving a highly populated area is enough to create a prolonged outage situation.

It is clear that a major space weather event causing significant disruption to the electrical grid in the United States could have paramount implications for the insurance industry. If businesses, utilities and households are without power for extended periods of time, insurers may be exposed to business interruptions.

The report caused a stir in the space weather community. The Swiss Academy of Sciences soon adopted its figures, adding that it would take between 4 and 10 years to repair.

Thus, the situation is becoming more serious. It would be necessary to estimate more precisely the probability that a Carrington event would occur again. A study, this time from England, claims that it is 1 every 100 to 200 years … Hmmmm… The last one happened 160 years ago. Is the next one imminent?

In truth, estimates differ. There are no reliable statistics since we only have 400 years of observations of the Sun. Others say that a Carrington-type event may occur only once every 500 years. But we are not much further ahead. These are only statistics, and the big flare may well occur … tomorrow!

What about larger solar storms? Are they possible? Japanese professor Kazunari Shibata has a wonderful idea. The ideal would be to be able to observe the Sun for say a thousand years. Difficult, isn't it? So why not observe a thousand solar-type stars for one year? Using data from NASA's Kepler satellite, and with the help of his first-year undergraduate students, he was able to observe 148 solar-type stars for 120 days. He found no less than 365 super-eruptions, stronger than the one that prevailed at the Carrington event. His conclusions are unequivocal: a solar storm a thousand times more energetic than the Carrington event can occur about every 5000 years. And a solar storm "only" one hundred times more energetic every 800 years. Such events are so violent that they occur over a very large area of the

Sun, making their probability of reaching the Earth very high. In fact, there is little chance for our planet to escape. Has this ever happened?

According to his colleagues Miyake, Nagaya, Masuda and Naka, the answer is yes. Such events produce a solar wind so fast that it can be called cosmic radiation: its protons have the same energy as those coming from distant stars. This radiation is detectable by its effects on Earth, particularly by the detection of carbon-14 and beryllium-10 in trees. Miyake and his colleagues have therefore analysed trunks that are several hundred years old. They found unambiguous increases in carbon-14 for several years after the year 775 AD, and then after 993. It seems that our Sun has already experienced, more than a thousand years ago, a storm of unprecedented magnitude. At that time, of course, neither radiation nor electricity was used…

No one knows the consequences of such a storm on our technological world. It is likely that no artificial satellite would survive, that astronauts would die, that air passengers would be exposed to fatal radiation, and no radio communication would be possible… Perhaps even worse: the core of nuclear power plants could melt. An end-of-the-world scenario. Should we fear it? And if, whatever we do, it is inevitable, what is the point of an ambitious space weather programme?

The answer comes from the experience of industrial disasters. Most of them do not have a single cause, but the conjunction of at least two causes. The 1989 magnetic storm that affected Quebec so much should have affected a much wider area. As it happened, the winter temperature in Quebec City was particularly low that day: the region had to deal with both severe weather and a solar event. In reality, there are almost always two or more causes for any industrial accident. The explosion of the AZF factory in 2001 in Toulouse had many more than two origins that are still being debated today. The Fukushima disaster in 2011 is a paroxysmal example: an earthquake, a tsunami, a power plant built in the wrong place…

Obsolescence of materials, disregard for safety rules, and poor training of personnel are often combined with an external event. The authorities in charge of safety are well-equipped to deal with one hazard: a flood, a heat or cold wave, torrential rains… They are much less equipped to deal with two simultaneous hazards. The solar weather hazard is not taken into account in civil security plans. What if, during a major volcanic eruption, a magnetic storm prevents communications from working? If lives need to be saved following an earthquake but planes are grounded as in Sweden in 2015 following a very small solar event? Space weather is not only the possible cause of a planetary disaster. It is, perhaps above all, an additional hazard which, added to other natural hazards, can lead to natural or industrial disasters. This is probably the reason why it is very important to develop forecasting capabilities.

Should we be afraid of it? Fear is a highly subjective and even cultural feeling. This is especially true in space weather.

The first definition we had at the very beginning of the 21st century was that of the US National Space Weather Plan: “Conditions on the Sun and in the solar wind, magnetosphere, ionosphere, and thermosphere that can affect the performance and reliability of space and ground-based technological systems and can endanger human life or health”. Besides the fact that this definition contains several words that are incomprehensible to the general public, it is decidedly fear-oriented. The aim is probably to convince the American senators to finance the research…

In 2007, researchers from 24 European countries proposed an alternative more in line with their approach, which is now the globally accepted definition. The UN World Meteorological Organization further improved it by clarifying the structure of this definition, which is as follows:

> *Space weather is:*
>
> *1. The physical and phenomenological state of the natural space environment, including the Sun and the interplanetary and planetary environments.*
>
> *2. The discipline that deals with the physical and phenomenological state of the natural space environment, including the Sun and the interplanetary and planetary environments. It aims, through observation, monitoring, analysis and modeling, at understanding the driving processes, predicting the state of the Sun, the solar wind, the magnetosphere, the ionosphere, the thermosphere and the Earth's magnetic field, monitoring and predicting their disturbances, and forecasting and nowcasting the potential impacts of these disturbances on ground-based or space-based infrastructure and human life or health.*
>
> *Space climate is the mean physical state of the space environment and its statistical variations in both space and time over a period of several solar cycles.*

These definitions banish fear. They replace it with the desire to know, to understand, to predict… Because among the most motivating arguments for studying space weather, there is science. To understand the universe, to dream in front of its beauty, to pierce its mysteries, for nothing, for free, without any objective, is a powerful motivation.

In the wonderful book "The Tartar steppe", by Dino Buzzati, the commander Giovanni Drogo waits in a fortress for the final and inevitably victorious battle that he will lead against an invisible enemy but always at the doors of the citadel. Space weather could look a bit like this. Since our industrial societies have existed, we have only had to deal with skirmishes. We know that a much larger solar flare will occur. We just don't know when. We also know that the more energy our societies consume, the more they depend on electricity, telecommunications, and transportation, and the more vulnerable they will be to this event that we are stalking with both anxiety and greed. The only solution, the wisest one, would be energy sobriety, the moderation of transport, and the economy of telecommunications. Humanity does not take this path. So we watch for an unpredictable enemy that seems never to come. Until the day when…

Acknowledgements

We thank Mr. Amsif Kadher for providing us with the CNES documents and giving us the space debris figures. We also thank Mr. Denis BOUSQUET, Global Chief Technical Officer and Space Underwriting at AXA XL for agreeing to communicate with us some of the figures on space insurance.

Chapter 6

Space Weather Operations

We have explored the history, physics, and applications of space weather and Space Climate. Now, it's time to delve into the core of operational forecasting. Previous chapters have covered the physics of Sun-Earth relations, and the instruments—both satellite-based and ground-based—that we use for observation. The challenge lies in how to arrange everything in a coherent picture. How do we utilise terabytes of daily data to extract clear, universally understandable, and immediately actionable insights? This dilemma only emerged in the mid-1990s and has literally exploded since then. We will attempt to address these questions. This chapter could represent an entire work on mathematics and statistics, computer science and its evolution, big data, and other contemporary tools, however, our work aims to chronicle the history of space weather and space climate. If at times it seems we stray too far from historical considerations, rest assured, we will swiftly return to them a few lines later! As we embark on this new chapter in the history of space weather and space climate, let's examine a class of key tools used in many fields of science, especially pertinent in operational prediction: indices.

Activity Index

When we are sick, we know it by several familiar symptoms: pain, fever, nausea. Yet, for an accurate diagnosis, doctors must assess our overall condition. What could be easier than a number to characterise it? The most common one is, of course, the body temperature, which serves as a reliable indicator of our state of health. However, temperature alone doesn't reveal the illness's cause, impact, or progression. It is even possible to be sick without a fever. And yet, its significance is undeniable.

In 1952, the obstetrical anaesthesiologist and medical researcher Dr. Virginia Apgar (1909–1974) saved the lives of hundreds of thousands of newborn infants by inventing the 10-point Apgar Score. She simply attributed 0–2 scores to a child's breathing, colour, reflexes, motion, and heart rate one minute after birth. This simple formula provided an objective and consistent assessment of the newborn's health and enabled the identification of babies in need of immediate support. Using standard indices, ratings, and scores alleviates the many frightening biases we humans are subject to, such as the time of our last meal, and so many others,

DOI: 10.1051/978-2-7598-3610-9.c006

as thoroughly explained by Daniel Kahneman in his book "Thinking, Fast and Slow", and enables us to make more accurate predictions than by just relying on subjective impressions of trained professionals.

Do we have similar classifications and indicators—in the field of space weather, we refer to them as indices—for solar and geomagnetic activity? Additionally, have we established any standard scales enabling us to make more accurate predictions?

Essentially, an index comprises a set of numerical values that provide information as relevant and reliable as possible on the phenomenon. Of course, the validity of an index primarily hinges on its precise definition and the specific aspect of the phenomenon it describes.

Observing the Sun from the Ground

In space weather, the oldest and perhaps most well-known index is the sunspot number. We have relied on it since chapter 1 to describe solar activity. It is the one that allowed the discovery of solar cycles.

However, it is more complicated than it seems. The first difficulty arises from the fact that to count the number of sunspots, you must be able to see them! We can only count those on the visible part of the Sun. Even then, sunspots are more easily observed in the centre of the solar disk than on its edges. Moreover, the quality of the observation instruments used in the 17th century bears no common measure with our modern instruments. Secondly, we are now able to detect micro-sunspots with very short lifetimes. However, we need continuity in our measurement: if we were to include these micro-sunspots that our predecessors could not observe, the number of detected spots would artificially increase. This would hinder our ability to compare the modern measurements with historical ones and ultimately determine the long-term trends of the Sun. Therefore, it is necessary to exclude these micro-sunspots from the final index, although we acknowledge their existence. A third challenge arises from the complex way in which Wolf defined the spot index in the mid-19th century. He combined the number of individual spots with the number of spot groups. This index is now called the International Sunspot Number and its acronym is the ISSN. Likely, we would not choose this approach today, but we are bound by the need to leverage long data sets. Due to this intricate definition, the sunspot number actually does not directly correlate with what an observer would see when projecting an image of the Sun through a telescope onto a wall. There are periods during which this index can reach values above 300 (during solar maximum) and periods with no sunspots at all (solar minimum).

Recently, an international team led by Belgian scientists re-evaluated the historical sunspot count. Frédéric Clette and Laure Lefèvre realised that if two observers were to draw a projection of the Sun—as was done until recently—it would result in two different numbers of sunspots. They delved into the notebooks of past experimenters and discovered many errors. For example, in 1947, a new counting practice emerged in which large spots were counted as 2, 3, or even up to 5 individual spots, depending on the size of their penumbra. Wolf himself attempted to align his observations with those of previous observers, such as Schwabe, introducing an observational bias. In 2014, their revisited sunspot series challenged many prior works. Such a sunspot deficit no longer had an astrophysical origin but simply a human one. Such a trend, widely commented on before, was only an artefact. The Maunder Minimum would not testify of a quieter Sun than the one experienced at the beginning of the 21st century. Their article caused, as one can imagine, quite a stir and one could, without exaggerating, describe it as a true "culture shock for the scientific community".

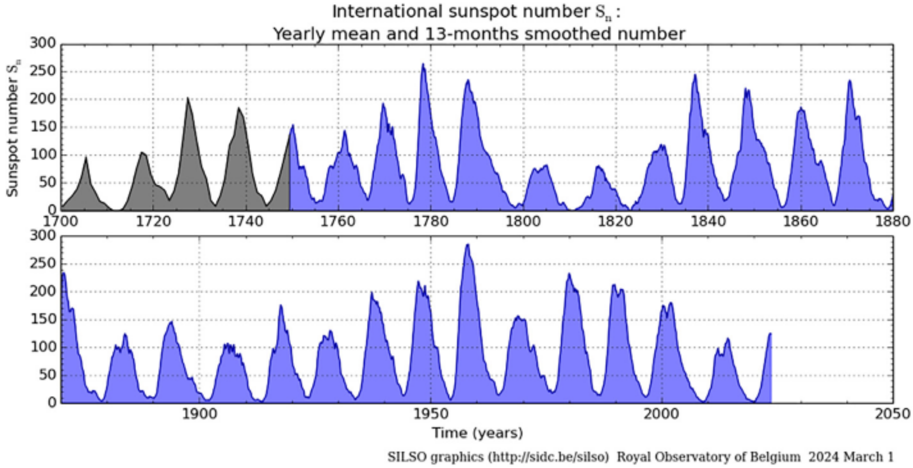


Fig. 6.1 – Yearly mean sunspot number (black) up to 1749 and monthly 13-month smoothed sunspot number (blue) from 1749 up to the present. Credit: WDC-SILSO, Royal Observatory of Belgium (https://www.sidc.be/SILSO/).

In addition to the indices defined to quantify space weather phenomena, space weather specialists also use different classification systems to organise and categorise observations. Sunspot regions, for example, have been observed since 800 BC by the Chinese and were even represented in manuscripts as early as the 12th century AD. Today, thanks to instruments such as a magnetograph capable of mapping the strength and location of magnetic fields on the Sun, sunspot regions are analysed in detail and classified into different categories based on their magnetic configuration and the number of sunspots they include. The categories are named after Greek letters (alpha, beta, gamma), which are sometimes appended with a suffix (gamma, delta). Alpha groups are unipolar, beta groups are bipolar, and gamma groups are multipolar. The gamma suffix is added when opposite polarities are irregularly distributed, and no straight line can be drawn between spots of opposite polarity. The delta suffix is added when the darkest regions of the sunspot (the umbrae) of opposite polarity are in a single penumbra, which is the outer, relatively lighter region on the spot. This classification system was developed by the Mount Wilson Observatory in California. Alpha and beta represent the simplest configurations, while beta-delta and beta-gamma-delta are more complex and considerably more prone to emit flares.

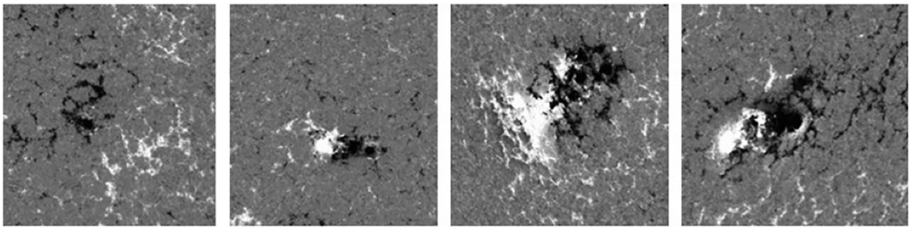


FIG. 6.2 – Images from a magnetogram showing examples of sunspot groups with alpha, beta, beta-gamma, and beta-gamma-delta configurations, from left to right.

Sunspots and sunspot groups are classified, measured, and used daily, but are not a panacea. Solar events can occur even without a sunspot being present. What else can be measured?

After the Second World War, the use of radars became widespread. Taking advantage of military surpluses, astronomers began their first observations of the universe in the radio domain. In Canada, Arthur Covington (1913–2001) and his colleagues at the National Research Council in Ottawa were equipped with a radar operating at 2800 MHz, corresponding to a wavelength of 10.7 cm. This wavelength was calculated by dividing the speed of light by the frequency. When they pointed the radar at the Sun, they were surprised to find that solar radiation at this wavelength varied. At that time, it was still unknown that the Sun could emit radio waves. Covington was able to demonstrate that this flux is linked to sunspots by taking advantage of a solar eclipse on November 23rd, 1946. As expected, there was a gradual decrease in the measured flux as the eclipse progressed over the Sun. However, each time the eclipse reached a sunspot, the drop was suddenly steeper, indicating that sunspots are major sources of this radiation. Since Covington's discovery, this measure has been used as a proxy for solar activity. It is called the flux at 10.7 cm (even if, strictly speaking, it is a flux density), or the decimetric index, noted f10.7. This index is expressed in 10^{-22} Wm^{-2} Hz^{-1}, a unit known as the solar flux unit. Unlike the number of sunspots, the value of the f10.7 index does not depend on the quality of the observer or the instrument. However, it is modulated by the changing distance between the Earth and the Sun due to the elliptical shape of the Earth's orbit. This modulation is corrected in the adjusted f10.7 index. The adjusted f10.7 index is still used today, although other wavelengths in the radio domain (around 3 cm) would have been more relevant. This value of 10.7 cm is therefore essentially due to chance! It has been recorded daily for decades. Canada is the international guarantor of this measurement. The f10.7 index ranges from around 70 for a quiet Sun to over 200 during an active period. The average of f10.7 over three months also provides an idea of the Sun's average activity. This value is used, in addition to the instantaneous value of the index, in atmospheric models.

We previously described how studying solar emission at 2800 MHz was useful to space weather and space climate scientists. But what about other frequencies? By examining radio emissions across the entire frequency spectrum using a

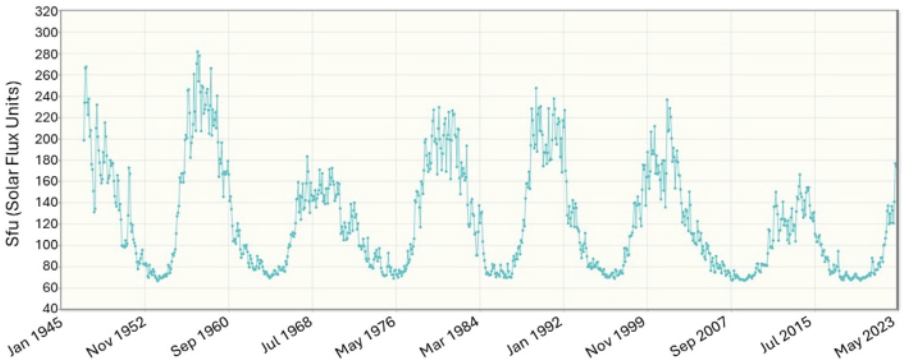


FIG. 6.3 – Averaged monthly Radio 10.7 cm flux (Adjusted) data. Credit: STAFF, Royal Observatory of Belgium.

spectrograph, we can analyse solar radio emissions and assess their impact on radio communication. In chapter 5, we highlighted the importance of this impact on the aviation sector, exemplified by the radio blackout that occurred on November 4, 2015. The spectrogram provides a visual representation in which the horizontal axis represents time, the vertical axis represents frequency, and the z-axis (or colour code) indicates the intensity of the electromagnetic radiation. Solar flares and coronal mass ejections manifest as radio bursts, clearly identifiable on the spectrogram. These radio bursts are classified into Types I, II, III, IV, or V based on the impacted frequencies, the drift over time, and the rate of the drift. For space weather, the most relevant types are Types II, III, and IV. Type III events (fast drift rate) are associated with solar flares, while Type II (slow drift rate) and Type IV (broadband long-duration burst) can indicate that a coronal mass ejection (CME) has taken place. Type II events are also often associated with radio blackouts and polar cap absorption events (which we will explain shortly). Notably, the absence of signals at low frequencies on a spectrogram can indicate a short-wave fadeout.

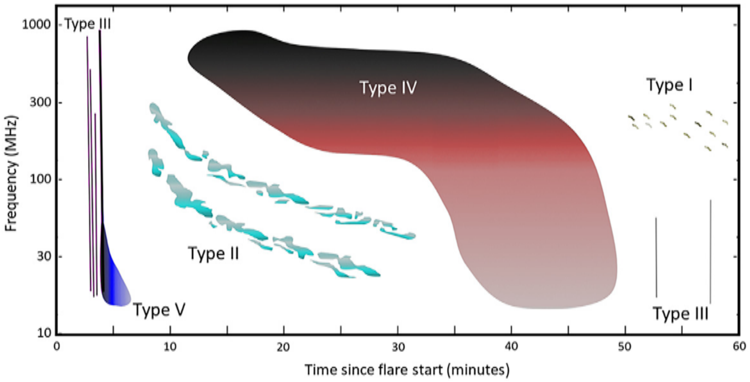


FIG. 6.4 – Radio spectrogram illustrating the different classes of radio bursts.

In addition, radio communication can also be degraded by the energetic particles accelerated during a flare. These particles get deflected by the Earth's magnetic field towards the poles, where they can degrade High Frequency (HF) communications for transpolar circuits. This phenomenon is called Polar Cap Absorption (PCA). To quantify the amount of electromagnetic wave absorbed in the ionosphere, scientists developed a device called a riometer (for Relative Ionospheric Opacity Meter for Extra-Terrestrial Emissions of Radio noise) in the 1950s at the University of Alaska. The riometer provides an indication of the severity of the PCA in decibels (dB) for different radio frequencies, typically around 30 MHz.

Since the space age, radio communication has also been crucial for communicating with satellites that require radio signals to cross the ionosphere. In the 1950s, while studying radio signals from distant sources, John Ashworth Ratcliffe (1902–1987) realised that radio-frequency signal amplitude and/or phase could fluctuate due to ionospheric irregularities. This phenomenon is known as scintillation and can lead to inaccuracies in GPS navigation systems, as described in chapter 5. The severity of scintillation is quantified with the help of two indices: S_4 and P_{rms} (or Φ_{rms} or sigma-phi). S_4 is the amplitude scintillation index which measures the amount of signal amplitude variation. P_{rms} is the phase scintillation index which measures the fluctuation in the signal phase. The S_4 index has a notional upper limit of 1.0. Scintillation values above approximately 0.3 may cause interference with satellite navigation and communication, while values above 0.6 could have severe effects, including loss of lock on the satellite. Severe phase scintillation may cause a loss of phase lock, potentially impacting space-based radars, for example.

Another aspect of space weather and space climate is its influence on Earth's magnetic activity. For millennia, humans have relied on Earth's magnetism for navigation. The magnetic compass, used by navigators for centuries, played a crucial role. However, in recent centuries, it became evident that solar events could impact compass readings. This realisation sparked interest in understanding and monitoring fluctuations in Earth's magnetic field. Even a slight variation could prevent a ship, its crew, and cargo from reaching their destination.

Once again, we require indices to assess the extent of Earth's magnetic field variation. Depending on the geographical area of interest, multiple indices may be necessary. While measuring magnetic field variations is relatively straightforward and cost-effective, precautions are essential. One challenge arises because the field measured by a magnetometer combines the internal field (also known as the secular field), which varies very slowly in time, but varies across location, with the external field, the field that concerns us in space weather and space climate. To assess the external field, we must remove the influence of the internal field from the measurement. Another challenge involves the need for well-distributed observatories across the globe. Unfortunately, this is impossible, due to the vast ocean coverage, which constitutes more than two-thirds of Earth's surface.

The first attempts at quantifying magnetic field variation date to the early 20th century. With only 13 observatories (including 2 in the southern hemisphere), Julius Bartels (1899–1964) and his colleagues proposed an index for magnetic field variation in 1939. He named it K, derived from the German word Kennziffer (code number), along with a ‘planetary’ average denoted as Kp. Notably, during that era, computers had not yet been invented, leading to a scale definition for K that is challenging to use in modern times. A zero index corresponds to a completely calm magnetic state, where only the internal field is measured. During the height of a geomagnetic storm, the index rises to 9. To refine this classification, he introduces plus and minus signs. For instance, an index of 7+ is slightly lower than an index of 8−. Additionally, we must consider the uneven distribution of stations when providing the Kp index. Weighting the measurements based on their origin becomes necessary. Although not the most convenient index, it remains popular among amateur aurora hunters. The larger the Kp, the farther from the poles auroras occur. This index is updated every three hours, with variations such as L (local), S (southern hemisphere), or N (northern hemisphere).

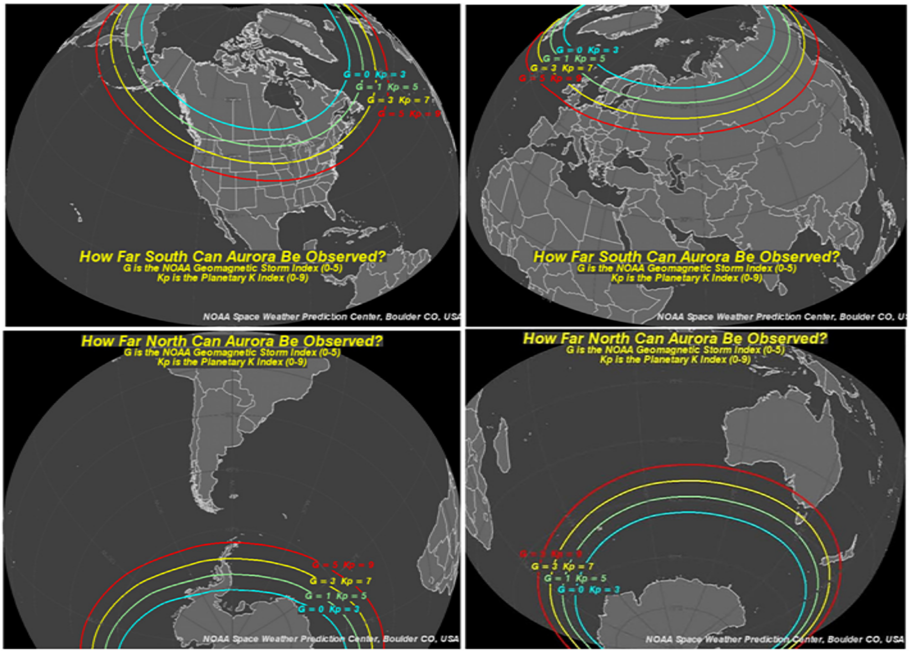


Fig. 6.5 – How far from the poles can an aurora be observed? Credit: NOAA Space Weather Prediction Centre.

With the advent of digital technology and the proliferation of observatories, other indices have emerged for research. In Bartels’ time, there were already 75 permanent magnetic observatories, with about a third located in Europe. Today, more than 150 observatories operate under programs like InterMagnet and the World Data Centre for Geomagnetism. We directly utilise the measurement of magnetic field variation, denoted as the index a (for equivalent amplitude), along with specific declinations: P (planetary), L (east–west), S (north–south), N (vertical, perpendicular to Earth’s surface), etc. The unit of measurement is the same as the magnetic field – nano Tesla.

Low activity is characterised by an Ap value close to zero. When calculated using historical K stations, it takes a capital letter: A.

In the early 1950s, as data increased, the scientific community realised that observatories distributed around the equator were particularly measuring the strength of the Earth's ring current. This current consists of charged particles in the equatorial plane of the Earth's magnetosphere at approximately 3–5 Earth radii, and it varies in response to changes in the solar wind. Taking only these near-equatorial observatories into account generated the Dst (Disturbance storm time) index. Dst is also of interest to auroral researchers and polar aurora enthusiasts because the ring current is not independent of the activity at high latitudes. At the equator, a geomagnetic storm is manifested by a decrease in the magnetic field. The more negative the Dst, the more energy is accumulated in the magnetosphere. A large magnetic storm is characterised by a Dst that can reach −250 nT and below.

Other groups of stations, this time in the auroral zone, provide researchers with specific indices. For example, the Polar Cap Index (PCI) developed in the 1960s, measures the level of geomagnetic activity in the polar regions of the Earth. It is used to track the intensity of high-latitude magnetic disturbances caused by solar wind particles interacting with Earth's magnetic field. All these indices share the same merits: continuity of measurement (they are provided every minute), reliability, and accessibility. They remain the foundation for any space meteorologist.

Observing the Sun from Space

Observing the Sun from space provides additional information not visible from the ground. Intense high-energy electromagnetic radiation, such as X-rays and extreme ultraviolet radiation emitted during solar flares, is absorbed by Earth's atmosphere and does not reach the Earth's surface. Since the 1980s, NASA has equipped its Geostationary Operational Environmental Satellite (GOES) with solar X-ray sensors that provide solar X-ray flux in Watts per square metre at different passbands. The peak X-ray flux at the 1–8 Angstrom passband is widely used for solar flare classification, as these wavelengths seem to best indicate the intensity of a flare. Solar flares are classified according to their strength. The smallest ones are A-class, followed by B, C, M, and X, the largest. The C, M, and X stand for Common, Medium, and eXtreme. In later years, as instruments became more sensitive, smaller flares became visible and were given the classes A and B. For example, a flux of 1×10^{-8} W/m^2 corresponds to an A1 flare, a flux of 2×10^{-8} W/m^2 corresponds to an A2 flare, and so forth until A9, then a flux of 1×10^{-7} W/m^2 corresponds to a B1 flare. The letter is changed every 10-fold increase. For fluxes greater than 1×10^{-4} W/m^2, the X letter is kept, and the digit can go above ten. The most powerful flare recorded so far was an X28 flare during the 2003 Halloween solar storms.

In addition to intense electromagnetic radiation, the Sun also emits high-energy charged particles consisting mainly of protons and electrons during a solar event. These particles can be accelerated to very high speeds and therefore have very high kinetic energies. The kinetic energy is measured in electronvolts (eV) and is related

Classification	Approximate peak flux range at 0.1–0.8 nm (watts/square metre)
A	$<10^{-7}$
B	10^{-7}–10^{-6}
C	10^{-6}–10^{-5}
M	10^{-5}–10^{-4}
X	$>10^{-4}$

to speed by the formula $E_c = ½\, mv^2$ where m is the mass, v is the velocity, and E_c the kinetic energy of the particle.

The GOES satellites are equipped with high-energy proton detectors that provide near real-time measurements of averaged integral proton flux. The sensor can measure the flux at different energy thresholds. The flux is expressed in protons per cm^2 per second per steradian (protons/cm^2/s/sr), where a steradian represents the unit of solid angle (a three-dimensional angle) in the International System of Units (SI). Typical solar proton energies range from a few hundred keV to hundreds of MeV. Given that the proton mass is 1.67×10^{-27} kg, these energies correspond to proton speeds ranging approximately from 5×10^6 m/s to 1.5×10^8 m/s. Electromagnetic radiation (such as visible light or X-rays), travels at the speed of light and, upon leaving the Sun, reaches the Earth in 8 min. In contrast, less energetic particles take a few hours to reach Earth, while the most energetic ones arrive in around 20 min.

In 1999, the US National Oceanic and Atmospheric Administration (NOAA) introduced Space Weather Scales[1] to describe the environmental disturbances related to three event types: geomagnetic storms (G-scale), solar radiation storms (S-scale), and radio blackouts (R-scale). These scales are based on the indices and measurements described above.

The tables below present the different scales, their descriptions, criteria, and the expected frequency of occurrence.

The S-scale is for solar radiation storms and is based on the proton flux at energies greater than 10 meV.

S-scale (solar radiation storms)	Description	Physical measure (flux level of particles at energies greater than 10 meV)	Average frequency (1 cycle = 11 years)
S 5	Extreme	10^5	Fewer than 1 per cycle
S 4	Severe	10^4	3 per cycle
S 3	Strong	10^3	10 per cycle
S 2	Moderate	10^2	25 per cycle
S 1	Minor	10	50 per cycle

---

[1] https://www.swpc.noaa.gov/noaa-scales-explanation.

The G-scale measures the severity of a geomagnetic storm and is based on the Kp index.

G-scale (geomagnetic storms)	Description	Physical measure	Average frequency (1 cycle = 11 years)
G 5	Extreme	Kp = 9	4 per cycle (4 days per cycle)
G 4	Severe	Kp = 8, including 9-	100 per cycle (60 days per cycle)
G 3	Strong	Kp = 7	200 per cycle (130 days per cycle)
G 2	Moderate	Kp = 6	600 per cycle (360 days per cycle)
G 1	Minor	Kp = 5	1700 per cycle (900 days per cycle)

The R-scale is a measure of the severity of Radio Blackouts (also known as shortwave fadeouts). It is based on the solar flare X-ray flux intensity.

R-scale (radio blackouts)	Description	Physical measure	Average frequency (1 cycle = 11 years)
R 5	Extreme	X20 (2×10^{-3})	Less than 1 per cycle
R 4	Severe	X10 (10^{-3})	8 per cycle (8 days per cycle)
R 3	Strong	X1 (10^{-4})	175 per cycle (140 days per cycle)
R 2	Moderate	M5 (5×10^{-5})	350 per cycle (300 days per cycle)
R 1	Minor	M1 (10^{-5})	2000 per cycle (950 days per cycle)

To illustrate all these indices, let's take a closer look at the late February 2023 event. On the 24th and 25th, two flares of class M erupted (visible on the GOES X-ray flux plot in figure 6.6). These flares were associated with radio bursts (visible in figure 6.7). Additionally, particle acceleration caused the proton flux to rise mildly (as can be seen in the GOES proton flux plot in figure 6.8). These protons led to a Polar Cap Absorption (PCA) in violet-coloured area in figure 6.9.

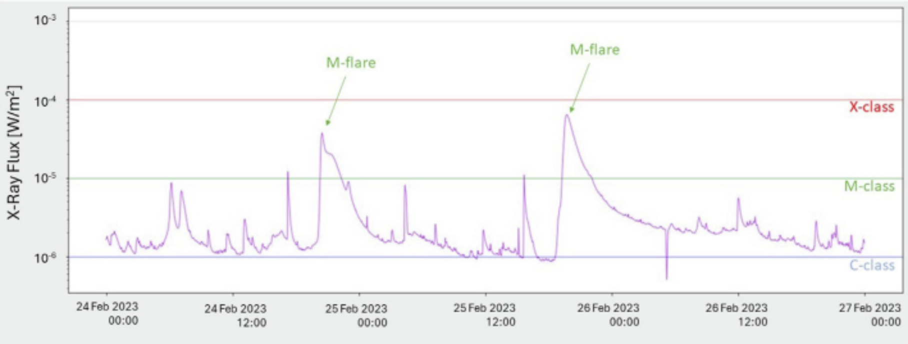


FIG. 6.6 – GOES X-ray flux in the 1–8 Angstrom passband showing the M flares on 24th and 25th of February 2023.

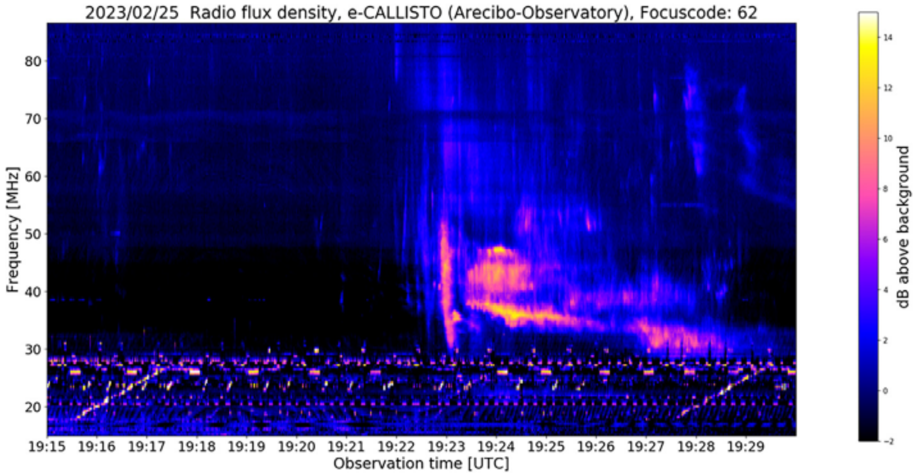


Fig. 6.7 – Radio Spectrogram from the e-callisto Puerto-Rico spectrographs. The orange patches indicated the 30–50 MHz radio waves emitted by the flares. Credit: Institute for Data Science FHNW Brugg/Windisch, Switzerland.

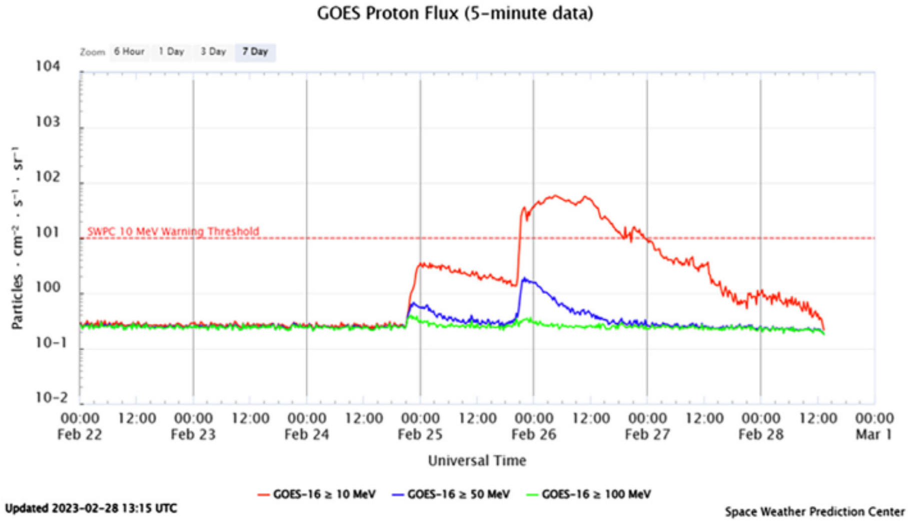


Fig. 6.8 – GOES proton flux showing the two separate rises due to the two separate flares. Credit: NOAA Space Weather Prediction Centre.

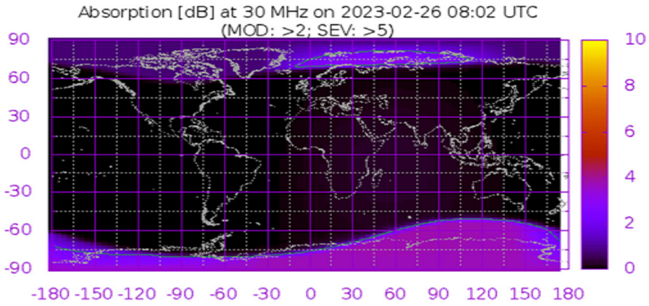


FIG. 6.9 – 30 MHz Radio signal absorption map. Credit: NOAA SWPC.

Furthermore, the flares were linked to two coronal mass ejections (CMEs). Upon reaching the Earth on the 26th and 27th of February 2023, these CMEs caused disturbances in the Earth's magnetic field. Specifically, the Dst index dropped down to −138 nT (figure 6.10) and the Kp rose above 6 (figure 6.11). This geomagnetic storm, classified as level G3 on the NOAA scale, generated beautiful auroras. These auroras were visible as far from the poles as parts of central Europe in the northern hemisphere and parts of southern Australia in the southern hemisphere.

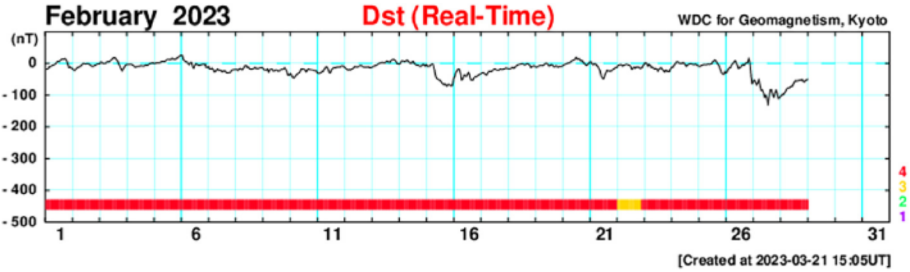


FIG. 6.10 – Dst index showing the drop below −100 from 26th of February 2023. Credit: WDC for Geomagnetism, Kyoto (https://wdc.kugi.kyoto-u.ac.jp).

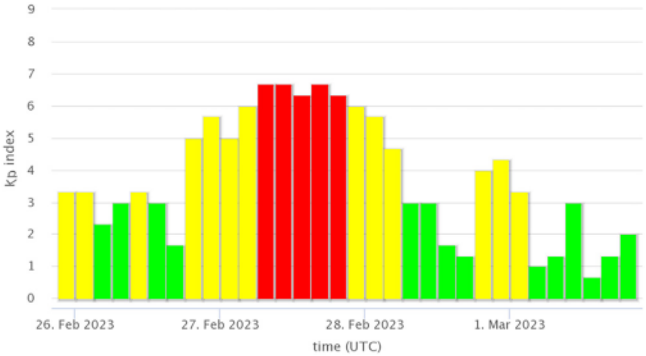


FIG. 6.11 – Planetary Kp index. Credit: GFZ German Research Centre for Geosciences.

Research on indices is an integral part of space meteorology. What information do these indices carry? Can we define more relevant indices? Should we consider as an index a physical quantity that is not directly measurable but deduced from several measurements, such as, for example, the energy flux entering the magnetosphere?

Just as body temperature alone doesn't reveal the nature of a disease, indices alone don't fully explain the underlying space weather or space climate phenomena. However, like a thermometer, indices serve as diagnostic tools, alerting us to deviations from normal conditions. They allow a simple description of massive data sets that vary with time (and possibly in space). They constitute the first quantification of a state of activity. To understand the symptoms, we must go further: we must dive into the incredible pile of measurements. This is the object of the rest of this chapter.

From Data to Forecast: The Key Role of Modelling

If you ask scientists what they do, most of the time their answer will be that, apart from writing or answering emails or writing funding applications, they acquire and analyse data. They try to understand what is behind it. They model the data by running curves through it with increasingly complex or even unconventional coefficients. They want to make sense of it to unlock the secrets of the world around us. But how to make sense of this data? This is where mathematics comes in. When we discover mathematics, we start with rather mundane basics: calculation rules, theorems and their reciprocals, and problems where we discuss a train's arrival time, or the quantity of melons bought according to the price of an apple. Very often, mathematics seems to be a potentially practical, but somewhat austere tool.

Yet it is the scientist's key to deciphering data. If you put yourself in the shoes of a person who wants to solve a problem, there is a whole protocol to implement to address it, akin to a recipe. First, we must pose the problem. What is the question we want to solve or, metaphorically speaking, which recipe are we going to prepare? Making a chocolate cake does not require the same ingredients as a cassoulet (a renowned French dish)! Once a problem is clearly defined, we must consider the vast amount of available data, essentially, the list of ingredients. Questions arise: What is the nature of the data? Which mission produced it? Are there any gaps or missing data? Do we need to apply corrections due to defects in the measuring instruments? As we contemplate these questions, the list of ingredients can become complex. After explaining the problem and collecting and cleaning the data, we arrive at a delicate yet essential step: analysing the data—the heart of the recipe. But which tools should we use? This is where mathematics enters the scene.

Mathematics provides numerous methods, but it's crucial to step back and assess whether a method aligns with the available data. Metaphorically, if we had only a frying pan available to bake a cake, it wouldn't be ideal. Sure, we might manage to bake a cake in a pan. But it's certainly not the best approach. There are more suitable methods, such as using an oven or even a microwave. In the context of cake baking, the choice of tools seems obvious. However, when dealing with space weather and space climate, finding the right tool becomes more intricate.

Modelling to Interpret Observations or to Simulate the Unobservable

In nature, when a phenomenon occurs—such as an aurora—we observe the end product in a certain way. The effects of each physical process at work are difficult to separate because they are so intricately intertwined. However, with the advancement of computer science and digital simulation, we can now reproduce and dissect the causes and effects one by one, and even anticipate them.

In the realm of numerical modelling for space plasmas, several approaches exist. To understand the logic and limitations of each approach, one must have some notion of how to describe a plasma.

The first method, which may immediately come to mind, involves studying each charged particle individually and examining its interactions with neighbouring particles. This particle-by-particle description is highly accurate but, due to the large number of particles involved, it demands significant computing power. In the 1950s, Buneman and Dawson developed methods for partitioning space into small boxes or cells (typically ranging from 100 to 1000). This approach is known as the Particle-in-cell (PIC) method.

Another approach is to establish a statistical description for a large number of particles. This representation captures the probabilities of finding a particle with a specific velocity within a given region of space. Instead of treating particles individually, we consider the probabilities associated with their states. An equation governs the evolution of these distribution functions. In this case, we no longer treat the particles one by one but by the probability that they must be in such or such a state. An equation governs the evolution of the distribution functions: the

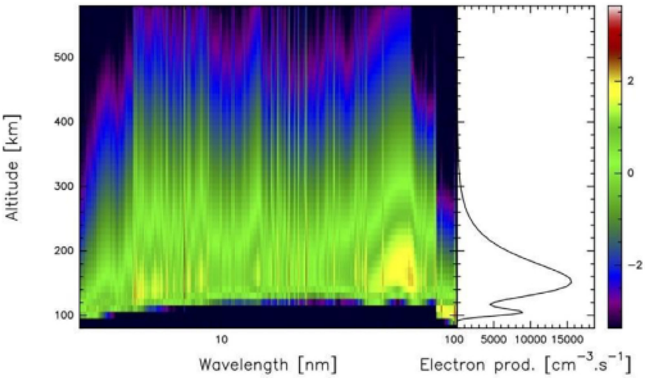


FIG. 6.12 – By resolving the Boltzmann equation, we can calculate, for instance, the altitude in the upper atmosphere where solar radiation deposits energy in the extreme ultraviolet, creating the ionospheric layers E and F. The left part of this figure illustrates the number of electrons produced at each wavelength. Yellow corresponds to 100 (= 10^2) electrons per cubic centimetre per second, while blue represents 0.01 (= 10^{-2}) electrons at the same rate. On the right, the curve represents their sum at all wavelengths. The main difficulty is that this is not reality, but only a model that represents it (Credit: Jean Lilensten, IPAG/CNRS).

Boltzmann equation. It balances the spatial and temporal variations of the distribution function with the sources and losses of particles. Depending on the assumptions made, this equation takes on different names. When dealing with plasmas and neglecting the binary interactions between particles, we refer to it as *the Vlasov equation*. This approach, called the kinetic approach, delves into the detailed evolution of the distribution functions for each category of particles (such as electrons and ions). It provides a comprehensive description of plasmas but comes with a significant computational cost.

A third and final method, faster to implement, is the magnetohydrodynamic (MHD) approach. In this approach, plasma is treated as a single magnetised fluid, and its properties are described using macroscopic quantities. These quantities include the density (measured in particles per unit volume), ensemble velocity, and temperature. We refer to these properties as moments (or mean values) of the distribution function. It is no longer a question of describing the evolution of a distribution function but of its moments. The MHD approach closely resembles classical hydrodynamics, with one key difference: it also incorporates equations describing the electric and magnetic fields. This method suffices for describing non-collisional or weakly collisional plasmas such as the solar wind.

To achieve a good level of plasma description while conserving computational resources, hybrid methods are sometimes employed. These hybrid approaches combine different modelling techniques. For instance, electrons may be treated using a fluid description (MHD), while ions are modelled kinetically.

In addition to modelling based on physical principles and equations, empirical models play a crucial role in describing the state of the magnetosphere or ionosphere statistically. Notably, the series of magnetospheric magnetic field models developed by Nikolai Tsyganenko relies on thousands of orbits of magnetic field measurements from onboard magnetometers on satellites. These models provide the three components of the magnetic field vector at any point in the magnetosphere, based on solar wind and interplanetary magnetic field conditions. On the ionosphere side, international teams, under the direction of Dieter Bilitza and Bodo Reinish, have established the IRI (International Reference Ionosphere) model. This model offers an altitude profile of the electron density for a specific location and time.

Recently, both physical and empirical models have been fed with data, sometimes in real-time, to achieve simulations that closely approximate reality. This technique is known as data assimilation. Additionally, artificial intelligence (AI) is playing an increasingly significant role in space sciences. Models can now be coupled with neural networks to generate forecasts. Machine learning algorithms analyse vast amounts of data from diverse sources, including satellites and ground-based instruments. By identifying patterns and correlations within the data, we enhance our ability to predict and mitigate the impacts of space weather and space climate events.

Correlation and Causality

Now, let's delve into a tool commonly used by the scientific community, one that can be both valuable and deceptive: correlation. When we aim to relate two temporally linked datasets, correlation appears as the most straightforward solution. We plot the evolution of one dataset against the other and assess the degree of their relationship. To illustrate further, consider the original example depicted in the figure below: an analysis of the correlation between annual per capita mozzarella consumption and the number of civil engineering graduates.

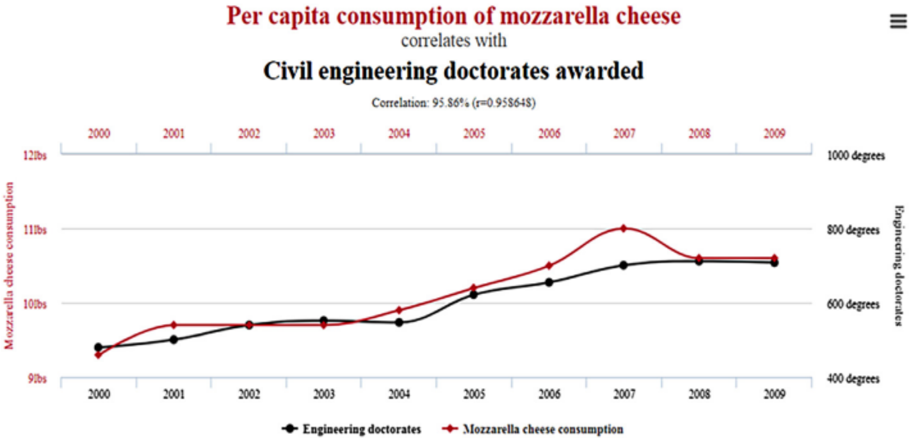


Fig. 6.13 – Retrieved from Tyler Vigen's website (https://www.tylervigen.com/spurious-correlations)—spurious correlations. This figure represents the temporal variation of the annual per capita mozzarella consumption and the number of civil engineering graduates.

The data seem to follow similar trends: the correlation coefficient is remarkably high at 95.56%! If the relationship between the nature of the data did not seem absurd, we would conclude that there is an obvious connection between them. However, in the context of space weather and space climate, it becomes much more complex to determine whether data can be correlated with each other. Scientists have been grappling with this challenge for decades and continue to do so today. Consider, for instance, attempting to correlate the variation of solar wind with the variation of the magnetic index Kp. Is it related to the square of the speed? Or perhaps the product of the velocity and the interplanetary magnetic field? The mechanisms governing Sun-Earth interaction are so intricate that the answer remains far from obvious. To address this complexity, we explore alternative methods. For example, we employ a sensitivity analysis using the Sobol index. With this approach, we study the influence of input uncertainty on the output. Imagine it as akin to assessing how the quality of eggs affects the softness of a chocolate cake. Quantifying this influence is challenging, but the Sobol index allows us to evaluate, for instance, whether the variation in solar wind speed has a greater impact on the magnetic index than the variation of the proton density in the solar wind. There are numerous other methods, and it's easy to get lost in this intricate landscape.

At this juncture, we must return to the source: what is the nature of our problem? When it comes to data analysis, we encounter two main categories: mathematical regression and classification. It is essential to recognise that mathematical regression holds a distinct meaning compared to common regression techniques. It is not, as it would be in everyday life, a step backwards, but rather an approximation of a variable. This variable, often difficult to measure, is approximated by another variable that is easier to access and that behaves similarly—thus it is correlated to the original. The variable closely related to the one we seek is sometimes referred to as a 'proxy'. Interestingly, in the context of space weather and space climate, the search for a suitable proxy is a common activity. The word classification is probably more familiar. In our understanding, however, classification is most generally done on statistical data.

In the context of the field that concerns us, scientists have oscillated between the regression problem and the classification problem, depending on the nature of the data to be predicted: do we want to predict an exact value of the magnetic index, *i.e.* a continuous value, or rather a level of disturbance of the magnetosphere with an interval of values? Once again, there is no perfect answer. Everything depends on the problem that has been posed and the data considered.

The Evolution of the Modelling of the Sun-Earth Interaction

Depending on our predictive goals, various methods are available, each adapted to the dynamics of the forecast. For instance, the radio emission flux originating from the Sun, for example, does not have the same behaviour as a geomagnetic index measured on the ground. Predicting these phenomena necessitates an understanding of how parameters vary over extended periods. Consequently, scientists have engaged in this long-term endeavour for nearly 50 years, establishing equations that connect solar wind parameters to magnetospheric activity resulting from solar interactions.

In the early 1970s, Robert McPherron[2] demonstrated that the magnetosphere behaves like an electrical capacitor, charging and discharging in response to energy accumulated during solar events. This initial study marked the beginning of a series of models aimed at connecting upstream and downstream processes across the more or less permeable barrier of the magnetosphere. Subsequently, scientists developed coupling functions to assess energy dispersion from the magnetosphere's entrance through various current systems. In 1978, Paul Perreault and Syun Ichi Akasofu proposed the first coupling function for analysing geomagnetic storms in terms of

---

[2] Numerous coupling functions have emerged, including that of Vytenis Vasyliunas, who employed dimensional analysis based on the physics of power extracted from the solar wind. His approach considered the relative importance of electromagnetic coupling, ionospheric conductivity effects, and viscous coupling within the magnetosphere. The list of researchers contributing to this field is extensive. For the curious reader, we can delve into the coupling functions proposed by Takashi Murayama in 1982, Lee Bargatze in 1986, Wen-Yao Xu and En-Qui Shi in the same year, Richard Stamper in 1999, and Ivan Finch and Mike Lockwood in 2007. Additionally, in 2007, Patrick Newell and his team aimed to create an almost universal coupling function by utilising magnetic indices and data from GOES or SuperDARN, correlated with various solar wind parameters. However, most energy coupling functions do not provide quantitative input in terms of energy from the solar wind due to coefficients that remain undetermined.

energy flow. They extracted solar wind parameters that control magnetic storms and their evolution. Their approach draws an analogy between the interplanetary energy flow and the electromagnetic power crossing a surface[3] (measured in W/m^2). They investigated the total energy dissipation rate in the magnetosphere in terms of injection of particles in the ring current, Joule dissipation due to ionospheric heating, and auroral particle injections. Notably, the correlation between geomagnetic and interplanetary parameters remains high, except during phases after the storm peak when the magnetosphere recovers and finds its equilibrium. Thereafter, scientists increased the complexity of the functions by considering the effects present within the magnetosphere to be more representative of its intricate mechanisms.

These analyses allow us to qualitatively study the relationships between solar wind parameters and magnetic indices.

However, in an operational forecasting framework that requires access to a proxy estimate with an evaluation of the associated value, these coupling functions cannot be used directly. In the past, other techniques aimed to study the behaviour of the magnetosphere in response to solar wind dynamics. For instance, in 1985, Bargatze and his team used linear filtering to obtain a generalised linear relation between information from the solar wind and that from the magnetosphere. This study confirmed two categories of response: a fast response associated with high activities, where magnetospheric activity is directly related to coupling with the solar wind, and a slower response associated with mechanisms internal to the magnetosphere. Although these findings remain qualitative, they play an important role in the development of operational models.

These works underscore the complexity of modelling the magnetosphere. Empirical relations provided by various analyses (such as the correlation of solar wind parameters with magnetic and magnetohydrodynamic indices, assimilation to the Poynting flux, and others) highlight the importance of certain solar wind parameters and help us better understand the impact of solar activity on different magnetospheric current systems. The nonlinearity of the magnetosphere's response to solar activity is also a key element in analysing the Sun-Earth interaction. To study this non-linearity, scientists have turned to models that directly implement it, providing new insights into our understanding of this interaction. These models also serve as operational tools for predicting the impact of solar activity on our magnetic environment and, consequently, our technologies.

A New Approach: Artificial Intelligence

In the context of space weather and space climate, we encounter the use of neural networks, machine learning, and the field of artificial intelligence.

Over the past three decades, machine learning has rapidly evolved and found applications in diverse domains, including space weather and space climate. This interest began with successful applications of this powerful technique to various problems, spanning fields such as finance, medicine, industrial production, geology,

---

[3] The physicist reader may guess without any doubt that it's the Poynting flux.

and physics. One of the key features of neural networks is their ability to learn from databases, allowing users to model data and establish precise rules that reveal underlying relationships between different attributes. Neural networks remain the most widely used method for providing operational magnetic index forecasts.

But what exactly are neural networks? To answer this question, let's rewind to the 1930s, specifically when Alan Turing (1912–1954) was a PhD student. During his thesis, this brilliant young mathematician sought to address a question posed by his mentor, David Hilbert (1862–1943): the problem of decidability, or more succinctly, 'Is there a solution for any logical problem posed?'

In his 1937 paper, Turing introduced a groundbreaking concept: the Turing machine. This theoretical construct laid the groundwork for all current computational systems, essentially serving as the first universal programmable computer. Turing's contributions extended beyond theory. His work during World War II, particularly in cryptography to decipher the Nazi Enigma machine, is well-documented. Historians estimate that his efforts shortened the war by one or two years and saved countless lives. To this day, Turing remains a revered figure in the field of computer science.

Shortly afterward, Turing published an article in Mind in 1950 that caused a stir and raised the question of Artificial Intelligence: 'Can machines think?' The second sentence of this article is fundamental because it announces all the subtlety of this method: 'It should start with the definitions of the terms machine and think.' At that time, the machines used for programming were far from being the powerful tools we use today; technologies have evolved significantly in 70 years. Finally, thinking is a very complex philosophical notion, often discussed during scientific debates when we use artificial intelligence.

If we go back to Turing's article published in Mind, he introduced the test known as the Turing test. To illustrate it simply, imagine the following situation: you are in a closed room with a computer to send messages. In an adjacent room, there is another computer with an artificial intelligence programmed into it. In yet another room, there is another computer and a human person to answer you. You will send messages virtually without ever knowing if the response will come from the computer or the human. However, you will still have to evaluate the origin of the answer. Surprisingly, in about 30% of the cases, it would be impossible for you to determine the answer's origin. This is because the computer, through an iterative process, gradually acquires a sufficient database to formulate a coherent response to the sender. How is all this possible? It's time to dive a little deeper into the world of artificial intelligence, in order to get back to space weather and space climate. But first, let's clarify the terms we use when we talk about artificial intelligence. Many notions are associated with it: machine learning, neural networks, deep learning etc.... often misinterpreted nowadays.

First of all, artificial intelligence (AI) refers to the set of theories and techniques implemented to allow a machine to simulate intelligence. This involves computational neurobiology, which uses calculation methods inspired by the functioning of our brain. Next, let's delve into machine learning, a field of study within artificial intelligence. It focuses on the design, analysis, development, and implementation of methods that enable a machine to evolve through a systematic process. Some of the methods that attempt to imitate real intelligence include neural networks, computer vision, and deep learning.

We will focus on describing neural networks because they have been used since the 1990s for forecasting magnetic indices and are still employed today at the Space Weather Prediction Center (SWPC), the US operational center that we will introduce later in this chapter. Neural networks exemplify the best of computational neurobiology.

The first definition dates to 1943 when McCulloch and Pitts published an article describing the concept of a neural network as 'a single cell, living in a network of cells, receiving inputs and generating outputs that depend on these inputs.' At first glance, this definition seems straightforward: it merely defines an entity that provides outgoing information based on incoming information. The foundation lies in our understanding of a biological neuron, translated into a mathematical function.

Consider the incoming information in a neuron, also known as the dendrite. To translate this into mathematics, we simply define this incoming information as the variable X. But it's not that simple! There's a crucial piece of information we mustn't overlook when translating incoming data into a neuron—this key factor plays a vital role in our forecasting models. Imagine that various pieces of information are reaching you right now—what you're reading, the weather, perhaps hunger, and more. Yet, not all this information holds the same level of importance in your thinking. This is where the concept of information weight comes into play. The higher the weight associated with an input variable X, the more dominant that information becomes in the neuron's processing. This weighted information reaches the core of the neuron, which activates based on the stimuli provided by the dendrite.

Mathematically speaking, it is the role of the scientist developing the artificial neuron to pay attention to how incoming data will be processed by it. To achieve this, activation functions are used. These functions, which can be more or less complex, process the incoming information and provide a trend for the outgoing information.

For instance, consider two types of people: binary thinkers and moderate thinkers. If you are a binary thinker, everything is either all or nothing. You would resemble a sign function: equal to 0 when the input variable is negative and equal to 1 when the input variable is positive. It's like being either hot or cold, but never lukewarm! On the other hand, if you are more moderate, you would exhibit behaviour similar to the "hyperbolic tangent" function. This function allows information to flow progressively based on its input value. Once the information reaches the core of the neuron and is processed by the activation function, it is transmitted to the axon at the output level. This process creates what we call a perceptron or formal neuron.

The first model of a perceptron was presented by the American psychologist Frank Rosenblatt in 1958. To connect these tiny neurons and transform them into powerful neural networks, synapses must be used to link them. This is where the real challenge for mathematicians and scientists begins!

We have just defined the perceptron, which serves as the elementary building block for constructing networks. To create increasingly complex structures, let's draw an analogy with cooking—creating a neural network is akin to making lasagna (fun fact: in the Python programming language, one of the first libraries developed for programming neural networks was called Lasagne!). In our network, we will stack layers upon layers of neurons. These layers will form more intricate networks, allowing us to connect incoming information to outgoing information and implement the non-linearity that bridges this information.

The simplest model to create is the multilayer perceptron, shown in the figure below.

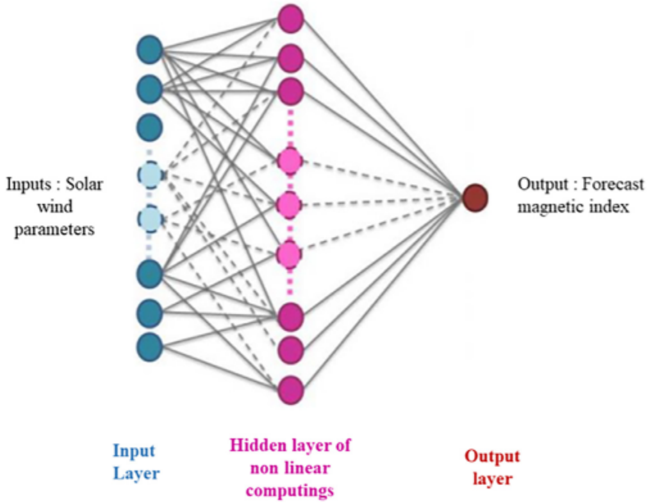


Fig. 6.14 – Illustration of the multilayer perceptron, the information goes from left to right through the different neurons illustrated by the coloured circles.

This is the first model that has been used in the context of space weather—an elementary model designed for all those who would like to delve into neural networks one day. The information enters a layer soberly called the input layer, where it is weighted. As we discussed earlier, the weight of the information is fundamental in determining the importance of one piece of information compared to another. The weighted incoming information is then analysed through layers of computation.

In space weather, scientists have always aimed to limit the number of computational layers to one because we face a physical problem. Increasing the number of hidden layers would raise the opacity of the calculations made internally by the network. The scientist's primary goal is to understand the link between solar wind parameters and the magnetic field indices to be forecast. A single layer allows us to establish non-linear computations while maintaining minimal control over the connections between incoming data and the outgoing data to be forecast. Finally, there is the output layer, which provides us with the forecast of the magnetic field index along with an error estimate—a topic that we will discuss shortly after presenting the models.

For the best-known models in the field, we can begin with that of Henrik Lundstedt and Peter Wintoft. In 1994, they utilised these networks to predict geomagnetic storms with the index Dst located at the equator. With this network and this initial simple approach, they demonstrated that it was possible to accurately predict the initial and main phases of a storm, but that it was more complex to forecast the recovery phase when the magnetosphere recovers. Shortly after, in 1997, Hans Gleisner and Lundstedt utilised this model once again to forecast the response of auroral electrojets to solar activity and observed how using non-linear functions could help enhance the relationships between solar wind and magnetic indices. In 2000, Fredrik Boberg and his team developed this network to provide real-time forecasts of the Kp index. More precisely, they developed a hybrid model to have a specific network for Kp predictions in calm periods and another for agitated periods. Indeed, achieving optimal performance in prediction with this type of network for each level of activity is complex, given the highly variable dynamics of the magnetosphere.

This network is probably too simple and does not provide very reliable information. Therefore, the scientists explored other more complex models of neural networks. However, the choice of the networks used was not arbitrary. Let's take the example of the Time Delay Neural Network. This network contains a time window illustrated in green in figure 6.15, which records what happened at the previous moment.

In other words, this means that the hidden layer of calculations receives the parameters of the solar wind at the present time, as well as from past instants, with a very specific weight related to temporality. You can see it coming; it is not insignificant! As we have emphasized in the last few paragraphs, the magnetosphere's response is immediate only when the solar event is extremely violent. Otherwise, the events accumulate gradually until they reach the limits of our magnetic shield's capacity, at which point it discharges like a capacitor.

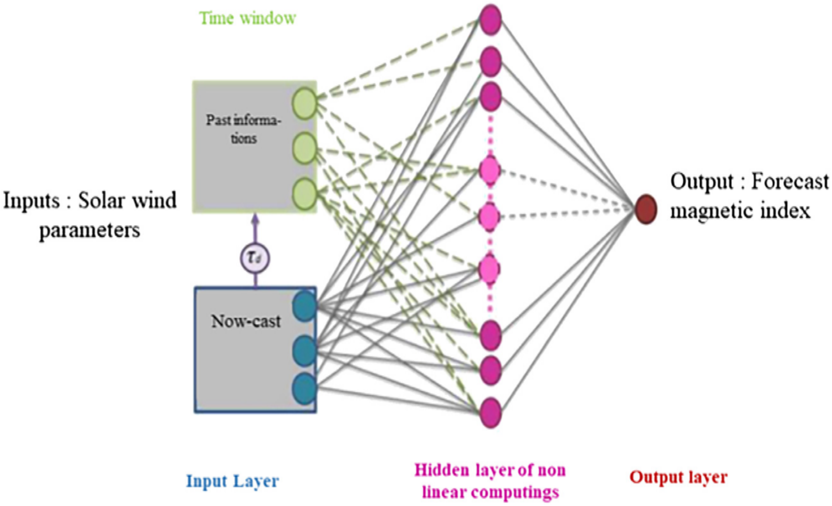


Fig. 6.15 – Illustration of the time delay network with the time window that records past information, connected to the input layer (credit from the authors).

Thanks to the time delay network, it is possible to consider this physical feature and create a better model for Sun-Earth interaction. In 1996, Gleisner used this network to predict the Dst index. Compared to the study conducted by Lundstedt and Wintoft in 1994, it was more efficient, providing a more accurate forecast for the recovery phase of a storm. Although this evolution may seem small, each time we observe progress through neural networks, it significantly impacts our understanding of the physical phenomena and *vice versa*.

Finally, another model emerged within the framework of magnetic index forecasting and is currently used for operational forecasts: the recurrent network. This network is a vast machine in which the hidden layers of calculations are influenced not only by incoming information (as was the case until now) but also by outgoing information. Why this evolution? The answer lies in a purely physical fact: the state of the magnetosphere at any given moment depends on its past state. If the magnetosphere has already been disturbed by a series of solar events, its reaction to a new event will differ from when it was initially calm. This consideration is crucial, as success entails achieving increasingly accurate predictions and a more refined understanding of magnetic indices.

In 2005, Wing and his collaborators developed three models based on the operation of these recurrent networks to predict the Kp index. These models (available on the NOAA website) rely on data provided by the ACE satellite, positioned at a stable point upstream of Earth. While these models offer real-time Kp predictions, they also underscore the complexity of studying the magnetosphere's response. Due to its internal dynamics, the magnetosphere either dominates the effects of external events or is itself dominated by them.

All these examples of neural networks and their multilayer, time delay, recursive, and other flavours represent just one type of machine learning algorithm, which is a subfield of artificial intelligence. Another subfield within artificial intelligence is deep learning, which has the capacity to recognise more complex patterns. However, this capability comes at the cost of requiring more sample data, more powerful hardware, and producing more opaque models. The challenge lies in understanding how deep learning algorithms arrive at their decision, making them less helpful in improving our understanding of physical processes.

Since Alan Turing, the field of artificial intelligence has made exponential progress. Today, it is widely used across various domains, including speech recognition, recommendation systems, self-driving cars, automatic decision-making, strategic game competitions, and even in the development of generative and creative tools. While artificial intelligence potentially presents ethical and moral challenges when integrated into everyday life, it also offers exciting opportunities to enhance our understanding, prediction, and mitigation of the negative impacts of space weather on critical infrastructure and technology.

At the Heart of Forecasting Models: The Data

It is time to focus on the heart of our problem: the data. Data, time and again, is the scientist's gold, yet it is so delicate to manipulate! So far, we have discussed tools for analysis and methods used to extract knowledge. However, let's take a moment to delve into the data itself. Doing so will reveal its complexity, especially when approached through artificial intelligence.

In the 21st century, a new profession has emerged: that of data analysis. A data scientist possesses three key skills: a programming background to code the models, a mathematical foundation (especially in statistics) essential for data analysis, and expertise in the relevant scientific field—in our case, space weather and space climate. Why is this profession increasingly prevalent today? Well, data is ubiquitous. We often say that each human being is datafied, meaning each of us represents a set of data. Our actions, preferences, movements, and purchases—all are meticulously recorded. Our world, and indeed our Universe, comprises petabytes of data. Over the past few decades, the volume of recorded data has exploded. Gone are the days of mere kilobytes; now we operate in an era where artificial intelligence is essential. We perform calculations in the cloud—sending our virtual computations to vast clusters of computers maintained by giants like Google and Amazon. These computational powerhouses accelerate processing times significantly. Yet, beyond data processing lies a new challenge: optimisation. We strive to make space weather forecasts within minutes, not hours, to keep them relevant. The data is precious, but it can also introduce bias. As we navigate this intricate landscape, we recognise that data holds immense potential and, simultaneously, the responsibility to wield it wisely. This is the main pitfall we aim to address, fundamentally associated with data processing: bias.

To illustrate the concept of bias, Google devised a straightforward game.[4] They asked a group of people to draw a shoe to determine if a certain shoe category stood out more than others. You can participate in this exercise too. Take 30 s and sketch the first shoe that comes to mind (you can do it in the box below).

« Draw me a shoe »

The most frequently returned model turned out to be sneakers, followed by heels, and to a lesser extent, city shoes. Now, let's consider using this image database of shoes to train a neural network. An intriguing phenomenon emerges: the algorithm excels at recognizing sneakers but struggles with other types of shoes. This phenomenon is what we call interaction bias. It arises from the way users influence the algorithm through their data choices during training.

FIG. 6.16 – Image from the Google video "Machine Learning and Human Bias".

To draw a culinary analogy, imagine you've had to practise making chocolatines, croissants and other sweet treats. Now, you're evaluated on your ability to prepare a cassoulet—a dish quite different from your prior learning. The evaluation wouldn't align with your training, even though both fall within the culinary domain. Scientists must be mindful of this bias to avoid introducing it into their analyses. For instance, in the context of space weather, data distribution analysis becomes crucial. When training an algorithm, should we exclusively use data from calm periods of solar activity? If we do, our algorithm might struggle to predict activity peaks. Achieving balance in the training data is essential. However, in the specific field of this work, reproducibility remains challenging. Two events are never perfectly identical. A geomagnetic storm can be linked to the subsequent peak of solar activity, such as coronal mass ejection. However, it can also result from an accumulation of energy. When this accumulated energy becomes too constrained, it breaks out and creates a storm. Here we delve into the complexity and beauty of solar event forecasting: analysing past events, establishing connections with physics to understand the intricate interactions, and evaluating the associated margin of error in our prediction.

The information provided by an algorithm, like a neural network, should not be regarded as exact and absolute science. Instead, it offers probabilistic information with an inherent error to assess. Often, a confusion matrix comes into play for evaluating prediction errors. From this matrix, several statistical indices emerge. For instance, the probability of detection and false alarm rate. The probability of detection informs the forecaster how many times a solar event has occurred and whether the algorithm successfully detected it (referred to as a true positive). The false alarm rate indicates how often a solar event was predicted but did not actually occur (termed a false positive). The algorithm might decide that no event is present. If there genuinely is no event, it's a true negative. However, if an event did occur, it is a false negative. When you examine forecasts provided by space weather forecasting sites, you won't receive 100% accurate information. Instead, it comes with a margin of error provided by the forecasters. One critical task for forecast tool designers is to optimise the algorithm's sensitivity (true positive rate) while maintaining sufficient specificity (true negative rate).

Forecasters, here we go! After discussing the importance of data, the evolution of models, the central role of physics and mathematics, and the numerous pitfalls of data analysis let's delve into what happens once an event occurs and is detected in an operational centre.

From Data to Forecast: The Operational Centres

Despite significant progress in observation, modelling and data processing in recent years, predicting solar events before they happen remains a challenge. To assess their potential impact, we still rely on observation satellites to detect these events. Recognising that these impacts result from a complex chain of interdependencies, we now refer to them as Sun-Earth interactions.

Progress hinges on the development of specialised models fuelled by consistent and relevant observations. These models rely on validation from the largest available dataset. However, the scarcity of observation points both on the ground and in space (given the vastness of our planet and the solar system), and our limited knowledge still impose significant constraints on our ability to create reliable forecasts. Nevertheless, ongoing satellite and ground data acquisition, along with the expansion of our historical database over time, contribute to continued progress. Active research, improved collaboration, and knowledge sharing also play a crucial role.

Given that space weather hazards are inherently global, international collaboration is essential for efficient and optimal response. Therefore, the International Space Environment Service (ISES) network was formed. It comprises 20 Regional Warning Centres (RWCs) that exchange real-time solar and geomagnetic activity data as well as forecasts. ISES has significantly benefited from two American initiatives. The first involves the establishment of a network of optical observatories—the Solar Observing Optical Network (SOON). Simultaneously, an identical network of radio monitoring stations—the Radio Solar Telescope Network (RSTN network)—was deployed on Air Force bases worldwide. The second initiative centres on placing American meteorological satellites into orbit. Specifically, the NOAA GOES satellites occupy geostationary orbits and are equipped with interplanetary X-ray and proton detectors. Additionally, X-ray data from the Japanese Yohkoh satellite was accessible to ISES several years before it became publicly available.

ISES facilitates collaborative efforts while safeguarding scientists' interests. Each forecasting centre within ISES focuses on specific areas of interest based on user needs. Consequently, each centre develops its unique methods and selects its specialised 'products'. The network enables detailed method comparison (including computer program evaluations) and validates shared forecasts. Despite cultural differences, each centre maintains its autonomy while fostering a cooperative environment.

FIG. 6.17 – The Royal Observatory of Belgium hosts one of the 20 regional warning centres distributed worldwide as part of the ISES network. Additionally, it participates in the Pan-European Consortium for Aviation Space Weather User Services (PECASUS network, https://pecasus.eu/), which is affiliated with the ICAO global aviation space weather network.

In parallel, several programs have emerged concerning Sun-Earth relationships and information sharing. Notably, the International Living with a Star (ILWS) and the International Space Weather Initiative (ISWI) serve as prominent examples, albeit primarily within the scientific domain. At the European level, the Space Safety Programme (S2P), led by the European Space Agency (ESA) aims to provide information and services related to the environment, threats (both human and natural) and sustainable space environment exploitation. Within the space weather segment of the Space Safety program, ESA is constructing the first European Space Weather Service Network. This network already provides an impressive amount of information to the general public and professionals. In alignment with Europe's diversity, the strategy involves leveraging existing capacities and federating them to deliver operational services. In the spirit of public service, all these products are freely accessible, even when they rely on complex and costly instruments, satellites, or intricate computer codes.

The governance of space-weather operational systems in Europe remains a topic of ongoing discussion. The European Commission, in conjunction with the establishment of its European Union Agency for the Space Programme (EUSPA), has delegated to ESA the management of research and innovation activities within the space weather domain. However, the budgetary allocation for this purpose is modest, standing at 1.2 million euros for 2023–24.

Since November 2019, the International Civil Aviation Organization (ICAO) has established a 24/7 global aviation space weather network. This network provides real-time and worldwide space weather updates for commercial and general aviation. The service generates and disseminates space weather advisories using the existing aeronautical fixed network for international aviation. Data is collected from dedicated space weather centres established by 17 countries, organised into four consortia: (1) ACFJ – Australia, Canada, France and Japan; (2) the Pan-European Consortium for Aviation Space Weather User Services (PECASUS) – Austria, Belgium, Cyprus, Finland, Germany, Italy, Netherlands, Poland, the United Kingdom and the South African National Space Agency; (3) the National and Oceanic and Atmospheric Administration (NOAA) Space Weather Prediction Centre (SWPC) operated by the United States, and (4) CRC – a centre operated by China and the Russian Federation. All global and regional centres focus on solar events that could potentially impact air transport-related high-frequency (HF) communications, GNSS-based navigation and surveillance, and radiation levels aboard civilian aircraft. This service directly provides space weather advisories to aircraft operators and flight crew members as part of their standard meteorological information relevant to their planned routes. These updates continue while they are in flight.

Numerous space weather centres operate worldwide, but one deserves special mention: the Space Weather Prediction Centre (SWPC). Located in Boulder (Colorado, USA), it is an element of the National Oceanic and Atmospheric Administration (NOAA) and its National Weather Service. The SWPC stands out as one of the most mature organisations in this field. It synthesizes and disseminates information about the space environment from a wide range of sources. Users can query the SWPC for past events, real-time data, or forecasts. Consequently, the SWPC increasingly serves as an expert resource for operators dealing with disturbances in the space environment. Notably, the SWPC benefits from both civilian (NOAA) and military (U.S. Air Force) contributors. The U.S. Air Force serves as the executive agency for space weather within the U.S. Department of Defence, providing operational forecasts in this domain.

In contrast, Europe lacks an integrated defence system, making it impossible to replicate the American model. Instead, Europe navigates between cooperation and national initiatives. Belgium led the way by establishing the first centre operational every day of the year, 24 h a day, back in 2006—the Solar Terrestrial Centre of Excellence (STCE).

Across Europe, many countries maintain their own centres, including the UK Met Office Space Weather Operations Centre, the Kanzelhöhe Observatory for Solar and Environmental Research in Austria, the Institute of Atmospheric Physics in the

Czech Republic, the Space Research Centre in Poland, the Norwegian Centre for Space Weather, the Finnish Meteorological Institute, the National Space Weather Service (SeNMEs) in Spain or the Alpine Space Weather Operations Centre in France, fully devoted to producing forecasts for the largest public throughout the world in their own languages…

A New Way Forward

Throughout history, as a new technology emerges, it happens that space weather reveals its vulnerability. Drawing from these past experiences, we must learn to consider space weather and space climate and fortify our technology against its adverse effects to the best of modern engineering's capability. However, equally crucial is the development of models and research that deepen our understanding of the Sun and enable predictions regarding its impact on our space environment. Additionally, we need to understand our own influence on both space and Earth environments. This includes addressing light pollution in the sky, managing the growing space debris population, and advancing instrument development and life cycles. Achieving these goals will necessitate international collaboration.

In April 2022, the European Space Weather and Space Climate Association (E-SWAN)—an international non-profit association, was established to unite, sustain, and advance space weather initiatives in Europe. E-SWAN's primary objectives include maintaining and enhancing international cooperation, fostering networks among scientists, engineers, stakeholders, and end-users, promoting space weather education, raising awareness and providing support to sectors potentially affected by space weather or space climate. Furthermore, E-SWAN represents the European space weather and space climate community in other global contexts.

The future of space holds excitement beyond our wildest dreams. Many more stories await, but for now, we conclude this chapter.

Bibliography

Akasofu S. I. (1981) Energy coupling between the solar wind and the magnetosphere, *Space Sci. Rev.* **28**, 121–190. https://doi.org/10.1007/BF00218810.

Bargatze L. F., Baker D. N., McPherron R. L., Hones E. W. (1985) Magnetospheric impulse response for many levels of geomagnetic activity, *J. Geophys. Res.: Space Phys.* **90**(A7), 6387–6394.

Boberg F., Wintoft P., Lundstedt H. (2000) Real time Kp predictions from solar wind data using neural networks, *Phys. Chem. Earth, Part C: Solar Terr. Planet. Sci.* **25**(4), 275–280.

Finch I., Lockwood M. (2007, March) Solar wind-magnetosphere coupling functions on timescales of 1 day to 1 year, *Ann. Geophys.* **25**(2), 495–506.

Gleisner H., Lundstedt H. (1997) Response of the auroral electrojets to the solar wind modeled with neural networks, *J. Geophys. Res.: Space Phys.* **102**(A7), 14269–14278.

Gleisner H., Lundstedt H., Wintoft P. (1996) Predicting geomagnetic storms from solar-wind data using time-delay neural networks, *Ann. Geophys.* **14**(7), 679.

Kahneman D. (2013, April 2) *Thinking, fast and slow, Farrar, Straus and Giroux*, 1st edn. ISBN: 978-0374533557.

Lundstedt H., Wintoft P. (1994) Prediction of geomagnetic storms from solar wind data with the use of a neural network, *Ann. Geophys.* **12**, 19–24.

Mayaud P. N. (1980) *Derivation, meaning, and use of geomagnetic indices.* Washington DC American Geophysical Union Geophysical Monograph Series, 22.

McCulloch W. S., Pitts W. (1943) A logical calculus of the ideas immanent in nervous activity, *Bull. Math. Biophys.* **5**(4), 115–133. [McPherron *et al.*, 1970] – McPherron R. L. (1970) Growth phase of magnetospheric substorms, *J. Geophys. Res.* **75**(28), 5592–5599.

Murayama T. (1982) Coupling function between solar wind parameters and geomagnetic indices, *Rev. Geophys.* **20**(3), 623–629.

Newell P. T., Sotirelis T., Liou K., Meng C. I., Rich F. J. (2007) A nearly universal solar wind-magnetosphere coupling function inferred from 10 magnetospheric state variables, *J. Geophys. Res.: Space Phys.* **112**(A1).

Perreault P., Akasofu S. I. (1978) A study of geomagnetic storms, *Geophys. J. R. Astron. Soc.* **54** (3), 547–573.

Rosenblatt F. (1958) The perceptron: A probabilistic model for information storage and organization in the brain, *Psychol. Rev.* **65**(6), 386.

Siegelmann H. T., Sontag E. D. (1991) Turing computability with neural nets, *Appl. Math. Lett.* **4** (6), 77–80.

Stamper R., Lockwood M., Wild M. N., Clark T. D. G. (1999) Solar causes of the long-term increase in geomagnetic activity, *J. Geophys. Res.: Space Phys.* **104**(A12), 28325–28342.

Turing A. M. (1937) On computable numbers, with an application to the Entscheidungsproblem, *Proc. London Math. Soc.* **2**(1), 230–265.

Turing A. M. (reedition 2009) Computing machinery and intelligence, *Parsing the Turing test.* Springer, Dordrecht, pp. 23–65.

Vasyliunas V. M., Kan J. R., Siscoe G. L., Akasofu S. I. (1982) Scaling relations governing magnetospheric energy-transfer, *Planet. Space Sci.* **30**(4), 359–365.

Wing S., Johnson J. R., Jen J., Meng C. I., Sibeck D. G., Bechtold K., Takahashi K. (2005) Kp forecast models, *J. Geophys. Res.: Space Phys.* **110**(A4). https://doi.org/10.1029/2004JA010500.

Wu J. G., Lundstedt H. (1996) Prediction of geomagnetic storms from solar wind data using Elman recurrent neural networks, *Geophys. Res. Lett.* **23**(4), 319–322.

Wu J.-G., Lundstedt H. (1997) Geomagnetic storm predictions from solar wind data with the use of dynamic neural networks, *J. Geophys. Res.* **102**, 14255–14268. https://doi.org/10.1029/97JA00975.

Xu W.-y., Shi E.-Q. (1986) Numerical examination of Akasofu's energy coupling function, *Chin. J. Space Sci.* **6**(1), 24–32.

Webography

Space weather forecasts: do your own!

This series of websites is not intended to be exhaustive. We would like to help you to become a space weather forecaster by browsing some sites in a logical series. While exploring the web, you will certainly find your own favourite sites, eventually in your own language. Remember also that they have a constant tendency to change their url's!

A general website
https://www.spaceweatherlive.com/

Look at the state of the Sun
SDO: https://sdo.gsfc.nasa.gov/data/dashboard/
SOHO: https://soho.nascom.nasa.gov/data/LATEST/current_c3.mp4
Classify the sunspots that you observed: https://www.stce.be/educational/classification

Solar flux (flares)
https://www.swpc.noaa.gov/products/goes-x-ray-flux
https://lasp.colorado.edu/eve/data_access/eve-one-minute-averages/index.html

General
https://www.sidc.be/spaceweatherservices/applications/solarmap/

Solar wind and arrival at Earth
https://www.sidc.be/spaceweatherservices/applications/cor2speed/cor2speed.html#canvas_position
flux at the subsolar point:
https://www.swpc.noaa.gov/products/ace-real-time-solar-wind
https://www.swpc.noaa.gov/products/real-time-solar-wind
A summary in plots: http://www.affects-fp7.eu/rssfeeds/ace_ap_forecast_plot/ace_realtime_ap_ch_gft_plot.png

Solar indices
Memento: https://sidc.be/educational/classification.php
https://www.swpc.noaa.gov/products/solar-cycle-progression
https://www.sidc.be/SILSO/home
https://omniweb.gsfc.nasa.gov/form/dx1.html

A complete computer code
https://helioviewer.org/

Look at the solar wind arriving at the Earth
Euphoria: https://swe.ssa.esa.int/current-space-weather
Enlil: https://www.spaceweatherlive.com/en/solar-activity/wsa-enlil.html

Look at the state of the Earth
https://www.sidc.be/SILSO/home
https://spaceweather2.uit.no/sw/
https://flux.phys.uit.no/Last24/
https://spaceweather.knmi.nl/viewer/

TEC: https://www.sws.bom.gov.au/Satellite/2/2
https://impc.dlr.de/products/total-electron-content/near-real-time-tec/near-real-time-tec-maps-global
Civil Aviation: https://www.ilmailusaa.fi/warnings.html#top=0#id=swx#select-area=4#FMILang=en
HF Blackouts: https://www.spaceweather.gov/news/; https://www.solarham.net/
Indices: https://www.swpc.noaa.gov/products/wing-kp

Compare to other forecasting centres
Note: you will find the list of the International Space Environment Service (ISES)
22 Regional Warning Centers, four Associate Warning Centers, and one Collaborative Expert Center at http://www.spaceweather.org/. Other forecasting centres include:
Australia: http://www.sws.bom.gov.au/Space_Weather
Belgium: https://www.sidc.be/#services
Canada: http://www.spaceweather.ca/index-en.php
China: http://eng.sepc.ac.cn/index.php
Czech republic: https://www1.asu.cas.cz/~sunwatch/weekly-forecast
France: http://www.meteo-espace.fr/fr; http://auroralpes.fr
Italy: http://eswua.ingv.it/ewphp/swpage/swpage.php; https://roma2.ingv.it/index.php/monitoraggio-e-sorveglianza/prodotti-del-monitoraggio/bollettini-di-space-weather?own=0 (Mediterranean era only)
Japan: http://swc.nict.go.jp/en/
Norway: https://spaceweather2.uit.no/sw/
Russia: http://forecast.izmiran.ru/en/index.php
South Africa: https://spaceweather.sansa.org.za
Spain: www.senmes.es
United States: https://www.swpc.noaa.gov/

ESA: https://swe.ssa.esa.int/current-space-weather
ISES: http://www.spaceweather.org/
NOAA: http://www.swpc.noaa.gov/communities/space-weather-enthusiasts
SOHO: http://sohowww.nascom.nasa.gov/spaceweather/
STCE: http://www.stce.be

Aurora maniac (Amateurs in French): https://www.aurora-maniacs.com/
Solar amateurs: http://www.solarham.net
Space weather live: https://www.spaceweatherlive.com/fr; http://www.spaceweather.org/
Sun synthesis: http://www.solen.info/solar/

Conclusion

We have in one breath – in one book – covered the recent and dazzling history of space weather and space climate. In a century, everything we thought we knew about the Sun and the Earth's space environment has shattered. The source of solar energy was thought to be chemical: it is nuclear; the Earth's atmosphere was thought to be confined to the first kilometres because of gravity: it stretches for hundreds of kilometres. There was a hint of discrete and sporadic relationships between the Sun and the Earth: they are permanent and can affect all parts of our technological society. In short, we have experienced a real conceptual evolution whose implications we have yet to fully comprehend. Some of its elements are part of the great revolution that science imposes on our conception of the Universe. It is, in our opinion, a revolution at least as important as the Copernican revolution, but few philosophers of science have yet seized it. Its paradigm is that the Universe is dynamic and often unstable at all time scales and all space scales. Space weather and space climate constitute a dazzling illustration.

But they have their own peculiarities. Thus, they reveal to mankind that we live in the solar atmosphere and that we are dependent on its variability. We have only recently explored the fantastic implications of this tremendous discovery, even while we are trained in operational forecasting. Here again, we have to question the paths on which we have cleared space since the 1950s because operability requires a huge qualitative leap. We need data in real-time, in considerable numbers, ingested by artificial intelligence, feeding physical models. We must also think on a large scale, to deploy instruments on the ground that cover the entire planet and are able to communicate with each other, but also with satellites in very diverse orbits.

We cannot wait for these efforts to stabilise. Space weather and space climate continue to move frantically forward. What challenges do they face?

First of all, they have to convince the industry that it is subject to its vagaries. As we pointed out, this is a difficult task. What industry boss would spend fabulous amounts of money to protect themselves from a problem that may never happen? So we have to be pedagogical, we have to work with industry to find ways to improve resilience to space weather without an unbearable additional cost, and we have to show them how to read space weather forecasts and what decisions to make to respond to an alert.

To raise public awareness, we also need an educational effort for which we are poorly equipped. The most striking example is this: wherever you go in the world, you can read a weather map by opening a newspaper, even without understanding the language. To do this, meteorologists around the world had to converge on a common representation of depressions, pressures and temperatures, and determine the parameters that would speak best to you, wind speed, cold waves… Such a goal seems light years away from space weather forecasters. But we are moving fast.

We learn from our elders. The time may not be that far away when, next to the usual weather map, we will get the Sun's map and that of our space environment. The same is true of television newscasts. Videographers, ahead of national channels, are already trying it.

There are significant scientific challenges. We talked at length about defining the worst case in chapter five. We must also look for what, on the Sun, indicates that there will be an event, a burst of activity. We call them precursors. There are other points that need to be clarified. For example, the phenomenon of the polar aurora is much more complicated than the simple emptying of magnetospheric regions. But how can we understand their phenomenal dynamics? Finally, we need to be imaginative in order to better observe the space environment and not miss out on essential information. How can we benefit from the night radiation that is only beginning to be explored?

On all fronts of science, space weather and space climate are making great strides. So great that they already have ambitions beyond our planet; several groups are already working on the space weather and space climate of planets.

We also discussed in chapter five the difficulty of having different cultures work together. In reality, the challenge of internationalisation goes far beyond that. Indeed, modern armies are users of space weather and space climate. China, the USA, Great Britain, France, Russia, Japan, India… in reality, few countries have both a powerful army and a space agency. For them, the temptation is great to reserve the most crucial information for the military. Space weather and space climate scientists are therefore faced with ideological choices: if they work for the army, they will have access to important resources, but will have to stop publishing their findings if the army asks them to. This real problem is all the more important in the multipolar world, where several military powers claim global sovereignty. How can we collaborate between Chinese, Americans and Europeans? What future do we want for space weather and space climate? Do we want a science that will be blocked against the block, or do we prefer it open to the world?

Answers to these questions will depend on our maturity. Space weather and space climate, barely born, must already behave as adults. The challenge is probably not the survival of humanity. But space weather and space climate will undoubtedly contribute to the future of mankind.