Lecture Notes in Physics

The Editorial Policy for Edited Volumes

The series Lecture Notes in Physics reports new developments in physical research and teaching - quickly, informally, and at a high level. The type of material considered for publication includes monographs presenting original research or new angles in a classical field. The timeliness of a manuscript is more important than its form, which may be preliminary or tentative. Manuscripts should be reasonably self-contained. They will often present not only results of the author(s) but also related work by other people and will provide sufficient motivation, examples, and applications.

Acceptance

The manuscripts or a detailed description thereof should be submitted either to one of the series editors or to the managing editor. The proposal is then carefully refereed. A final decision concerning publication can often only be made on the basis of the complete manuscript, but otherwise the editors will try to make a preliminary decision as definite as they can on the basis of the available information.

Contractual Aspects

Authors receive jointly 30 complimentary copies of their book. No royalty is paid on Lecture Notes in Physics volumes. But authors are entitled to purchase directly from Springer other books from Springer (excluding Hager and Landolt-Börnstein) at a $33\frac{1}{3}$% discount off the list price. Resale of such copies or of free copies is not permitted. Commitment to publish is made by a letter of interest rather than by signing a formal contract. Springer secures the copyright for each volume.

Manuscript Submission

Manuscripts should be no less than 100 and preferably no more than 400 pages in length. Final manuscripts should be in English. They should include a table of contents and an informative introduction accessible also to readers not particularly familiar with the topic treated. Authors are free to use the material in other publications. However, if extensive use is made elsewhere, the publisher should be informed. As a special service, we offer free of charge LaTeX macro packages to format the text according to Springer's quality requirements. We strongly recommend authors to make use of this offer, as the result will be a book of considerably improved technical quality. The books are hardbound, and quality paper appropriate to the needs of the author(s) is used. Publication time is about ten weeks. More than twenty years of experience guarantee authors the best possible service.

LNP Homepage (springerlink.com)

On the LNP homepage you will find:
—The LNP online archive. It contains the full texts (PDF) of all volumes published since 2000. Abstracts, table of contents and prefaces are accessible free of charge to everyone. Information about the availability of printed volumes can be obtained.
—The subscription information. The online archive is free of charge to all subscribers of the printed volumes.
—The editorial contacts, with respect to both scientific and technical matters.
—The author's / editor's instructions.

K. Scherer H. Fichtner B. Heber U. Mall (Eds.)

Space Weather

The Physics Behind a Slogan

Springer

Editors

Klaus Scherer
Universität Bonn
Institut für Astrophysik
und Extraterrestrische Forschung
Auf dem Hügel 71
53121 Bonn
Germany

Bernd Heber
Universität Osnabrück
Fachbereich Physik
Institut für Experimentalphysik
Barbarastr. 7
49069 Osnabrück
Germany

Horst Fichtner
Universität Bochum
Institut für Theoretische Physik
Lehrstuhl IV:
Weltraum- und Astrophysik
44780 Bochum
Germany

Urs Mall
Max-Planck-Institut
für Sonnensystemforschung
Max-Planck-Str. 2
37191 Katlenburg-Lindau,
Germany

K. Scherer, H. Fichtner, B. Heber, U. Mall (Eds.), *Space Weather*, Lect. Notes Phys. **656** (Springer, Berlin Heidelberg 2005), DOI 10.1007/b100037

Library of Congress Control Number: 2004116344

ISSN 0075-8450
ISBN 3-540-22907-8 Springer Berlin Heidelberg New York

Typesetting: Camera-ready by the authors/editor
Data conversion: PTP-Berlin Protago-TEX-Production GmbH
Cover design: *design & production*, Heidelberg

Printed on acid-free paper
54/3141/ts - 5 4 3 2 1 0

Lecture Notes in Physics

For information about Vols. 1–610
please contact your bookseller or Springer
LNP Online archive: springerlink.com

Vol.611: A. Buchleitner, K. Hornberger (Eds.), Coherent Evolution in Noisy Environments.

Vol.612: L. Klein, (Ed.), Energy Conversion and Particle Acceleration in the Solar Corona.

Vol.613: K. Porsezian, V.C. Kuriakose (Eds.), Optical Solitons. Theoretical and Experimental Challenges.

Vol.614: E. Falgarone, T. Passot (Eds.), Turbulence and Magnetic Fields in Astrophysics.

Vol.615: J. Büchner, C.T. Dum, M. Scholer (Eds.), Space Plasma Simulation.

Vol.616: J. Trampetic, J. Wess (Eds.), Particle Physics in the New Millenium.

Vol.617: L. Fernández-Jambrina, L. M. González-Romero (Eds.), Current Trends in Relativistic Astrophysics, Theoretical, Numerical, Observational

Vol.618: M.D. Esposti, S. Graffi (Eds.), The Mathematical Aspects of Quantum Maps

Vol.619: H.M. Antia, A. Bhatnagar, P. Ulmschneider (Eds.), Lectures on Solar Physics

Vol.620: C. Fiolhais, F. Nogueira, M. Marques (Eds.), A Primer in Density Functional Theory

Vol.621: G. Rangarajan, M. Ding (Eds.), Processes with Long-Range Correlations

Vol.622: F. Benatti, R. Floreanini (Eds.), Irreversible Quantum Dynamics

Vol.623: M. Falcke, D. Malchow (Eds.), Understanding Calcium Dynamics, Experiments and Theory

Vol.624: T. Pöschel (Ed.), Granular Gas Dynamics

Vol.625: R. Pastor-Satorras, M. Rubi, A. Diaz-Guilera (Eds.), Statistical Mechanics of Complex Networks

Vol.626: G. Contopoulos, N. Voglis (Eds.), Galaxies and Chaos

Vol.627: S.G. Karshenboim, V.B. Smirnov (Eds.), Precision Physics of Simple Atomic Systems

Vol.628: R. Narayanan, D. Schwabe (Eds.), Interfacial Fluid Dynamics and Transport Processes

Vol.629: U.-G. Meißner, W. Plessas (Eds.), Lectures on Flavor Physics

Vol.630: T. Brandes, S. Kettemann (Eds.), Anderson Localization and Its Ramifications

Vol.631: D. J. W. Giulini, C. Kiefer, C. Lämmerzahl (Eds.), Quantum Gravity, From Theory to Experimental Search

Vol.632: A. M. Greco (Ed.), Direct and Inverse Methods in Nonlinear Evolution Equations

Vol.633: H.-T. Elze (Ed.), Decoherence and Entropy in Complex Systems, Based on Selected Lectures from DICE 2002

Vol.634: R. Haberlandt, D. Michel, A. Pöppl, R. Stannarius (Eds.), Molecules in Interaction with Surfaces and Interfaces

Vol.635: D. Alloin, W. Gieren (Eds.), Stellar Candles for the Extragalactic Distance Scale

Vol.636: R. Livi, A. Vulpiani (Eds.), The Kolmogorov Legacy in Physics, A Century of Turbulence and Complexity

Vol.637: I. Müller, P. Strehlow, Rubber and Rubber Balloons, Paradigms of Thermodynamics

Vol.638: Y. Kosmann-Schwarzbach, B. Grammaticos, K.M. Tamizhmani (Eds.), Integrability of Nonlinear Systems

Vol.639: G. Ripka, Dual Superconductor Models of Color Confinement

Vol.640: M. Karttunen, I. Vattulainen, A. Lukkarinen (Eds.), Novel Methods in Soft Matter Simulations

Vol.641: A. Lalazissis, P. Ring, D. Vretenar (Eds.), Extended Density Functionals in Nuclear Structure Physics

Vol.642: W. Hergert, A. Ernst, M. Däne (Eds.), Computational Materials Science

Vol.643: F. Strocchi, Symmetry Breaking

Vol.644: B. Grammaticos, Y. Kosmann-Schwarzbach, T. Tamizhmani (Eds.) Discrete Integrable Systems

Vol.645: U. Schollwöck, J. Richter, D.J.J. Farnell, R.F. Bishop (Eds.), Quantum Magnetism

Vol.646: N. Bretón, J. L. Cervantes-Cota, M. Salgado (Eds.), The Early Universe and Observational Cosmology

Vol.647: D. Blaschke, M. A. Ivanov, T. Mannel (Eds.), Heavy Quark Physics

Vol.648: S. G. Karshenboim, E. Peik (Eds.), Astrophysics, Clocks and Fundamental Constants

Vol.649: M. Paris, J. Rehacek (Eds.), Quantum State Estimation

Vol.650: E. Ben-Naim, H. Frauenfelder, Z. Toroczkai (Eds.), Complex Networks

Vol.651: J.S. Al-Khalili, E. Roeckl (Eds.), The Euroschool Lectures of Physics with Exotic Beams, Vol.I

Vol.652: J. Arias, M. Lozano (Eds.), Exotic Nuclear Physics

Vol.653: E. Papantonoupoulos (Ed.), The Physics of the Early Universe

Vol.654: G. Cassinelli, A. Levrero, E. de Vito, P. J. Lahti (Eds.), Theory and Appplication to the Galileo Group

Vol.655: M. Shillor, M. Sofonea, J.J. Telega, Models and Analysis of Quasistatic Contact

Vol.656: K. Scherer, H. Fichtner, B. Heber, U. Mall (Eds.), Space Weather

Preface

The topic of the second spring school organized by the "Arbeitsgemeinschaft Extraterrestrische Forschung"(AEF) under the auspices of the "Deutsche Physikalische Gesellschaft" (DPG) was the physics of space plasmas with special emphasis on the so-called solar-terrestrial relations. This traditional term for the various processes that connect the physics of the Sun with that of the Earth and its environment has been substituted to a large extent with the slogan "Space Weather" during recent years. Given that the physics of solar-terrestrial relations has borrowed many words from meteorology – like solar wind, magnetic clouds, polar rain – the term is certainly justified and doubtlessly beneficial as it helps to communicate the fascinating subject of space plasma physics to the scientific community as well as the public. A general interest in the topic is well documented by numerous scientific and non-scientific publications in all media. Consequently, it was a timely task to organize a school on "Space Weather" with the motivation to demonstrate to young physicists the value, the relevance as well as the beauty of space plasma physics.

At the present time of growing interest, however, one can observe with increasing frequency that the new term is often used as mere slogan empty of any physics or is misused in various contexts, up to a point where it may even become counter-productive, at least for scientific purposes. Therefore, already the full title of the school, "Space Weather – The Physics Behind a Slogan", was chosen to reveal its basic intention. To fulfil the main goals of the school, namely to explain what Space Weather refers to and to demonstrate how space plasma physics is connected to it as well as to basic science, a number of leading scientists in the field of solar-terrestrial relations were invited. This group was complemented by experts of related fields dealing with, e.g., hazards to manned space missions. Most of the speakers agreed to write up their lectures in detail. These lectures were, where it was possible, prepared in a concerted effort in order to avoid repetitions and form the following chapters. These are written rather in the style of a textbook than in the style of scientific articles. Such a textbook on Space Weather is still missing, and this volume of lecture notes contributes to fill this gap.

We thank all lecturers for taking the time to prepare, to present as well as to write up their lectures and to involve the participants of the school into many lively discussions.

Beyond the physics a lot of preparatory and organizational work had to be done. We are most grateful to Dr. Victor Gomer from the "Physikzentrum Bad Honnef" of the DPG. Dr. Gomer made the organizers lives easy and supported the school in a convenient unbureaucratic manner. We also acknowledge full support from the DPG and the "Wilhelm und Else Heräaus Stiftung" to the school that took place from March 30 to April 4, 2003.

Gillersheim,
September 2004

Klaus Scherer
Horst Fichtner
Bernd Heber
Urs Mall

Contents

Part III Models of Space Weather

Part IV Consequences of Space Weather

Introduction to Space Weather

Daniel N. Baker

Laboratory for Atmospheric and Space, University of Colorado, Boulder, CO 80303, USA

Abstract. Adverse space weather is one of the principal threats to modern human technology. Solar coronal mass ejections, large solar flares, and high-speed solar wind streams often lead to sequences of damaging disturbances within the Earth's magnetosphere, in the atmosphere, and even on the Earth's surface. Powerful and long-lasting geomagnetic storms can develop following solar disturbances and enhancements of the highly relativistic electron populations throughout the outer terrestrial radiation zone can also result. High-energy protons and heavier ions arriving in near-earth space – or trapped in the magnetosphere and having clearest effect in the South Atlantic Anomaly (SAA) – can damage satellite solar power panels, confuse optical trackers, and deposit harmful charges into sensitive electronic components. Recent international space science programs have made a concerted effort to study activity on the Sun, the propagation of energy bursts from the Sun to near-Earth space, energy coupling into the magnetosphere, and its redistribution and deposition in the upper and middle atmosphere. Extreme solar, geomagnetic and solar wind conditions can be observed by a large array of international satellites and ground-based sensors. Many types of space weather-related problems have been identified in recent years. This chapter presents examples of space weather-induced anomalies and failures and discusses community efforts to propose technical and operational solutions to space weather problems now and in the future.

1 Introduction

Above the thin layer of Earth's atmosphere where normal weather occurs (the troposphere), there is a vast region extending into interplanetary space that is permeated by highly fluctuating magnetic fields and very energetic particles. The collective, often violent, changes in the space environment surrounding the Earth are commonly referred to as "space weather". For several decades now, humans have increasingly used space-based assets for navigation, communication, military reconnaissance, and exploration. New observations, numerical simulations, and predictive models are helping to make important strides to deal with (if not alter) space weather (National Space Weather Program Strategic Plan, 1995 [13]; Baker, 1998 [2]).

As shown by Fig. 1 (taken from NASA "Roadmap" documents), the Sun and its interaction with the Earth is a prototype for much of our understanding of cosmic plasma physics. The upper chain of insets suggest that our understanding of the fundamental elements of magnetospheric physics,

D.N. Baker, Introduction to Space Weather, Lect. Notes Phys. **656**, 3–20 (2005)
http://www.springerlink.com/

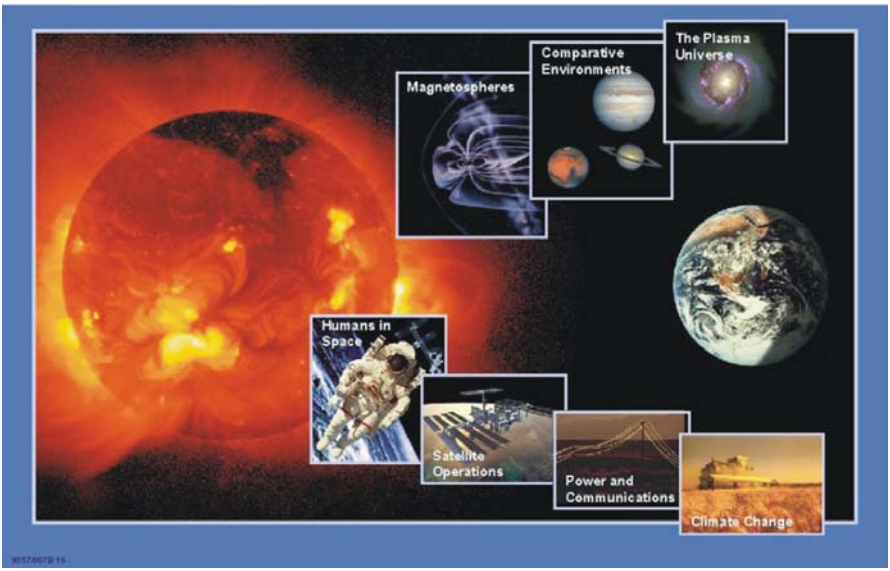


Fig. 1. Scientific and applications-related aspects of the Sun-Earth Connections research (courtesy of NASA).

our approach to comparative planetary environments, and ultimately our understanding of the plasma universe springs from our studies of Sun-Earth connections.

The lower chain of insets in Fig. 1, however, makes another important point: The space environment that we study for its intrinsic science value (as just described) is also an environment that has crucial practical importance. The effects of the space environment on humans in space, spacecraft operations, communications systems, power systems, and even (possibly) on climate make the understanding of Sun-Earth connections a manifestly important subject from a very pragmatic standpoint. Thus, the space weather "branch" of Fig. 1 is highly important much as is the basic science "branch".

Figure 2 shows in a schematic way the linked Sun-Earth system. It is known from several decades of research that the Sun is the overwhelming driver of space weather effects in near-Earth space. The solar wind emanating from the Sun – and the embedded interplanetary magnetic field (IMF) – provides the momentum, the energy, and much of the mass that fills and powers the Earth's magnetosphere. The Earth's ionosphere and atmosphere responds to this solar wind driving in complex ways. The ionosphere can also supply particles (mass) to populate the terrestrial magnetosphere and, of course, the neutral atmosphere responds strongly to solar irradiance (photons) as well as to plasma interactions with the solar wind.

The magnetosphere-ionosphere-atmosphere system is immensely complicated and constitutes a high-coupled system (see Fig. 3). There are several

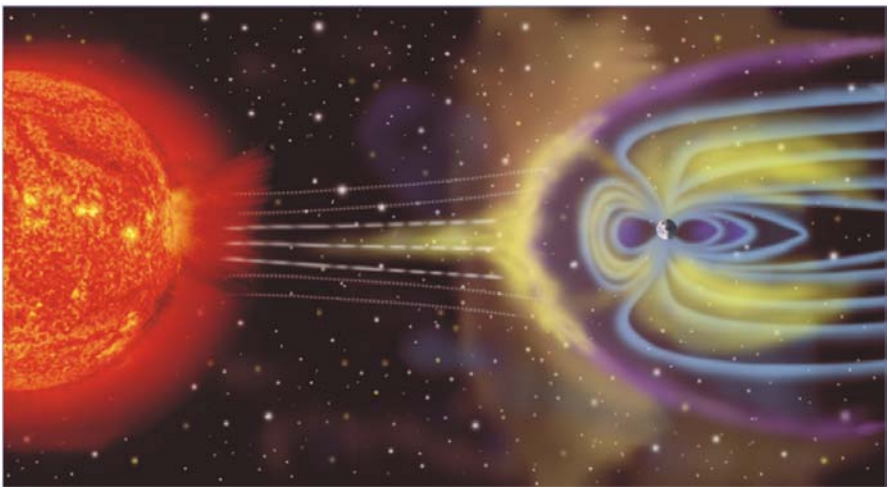


Fig. 2. The Sun, the interplanetary medium, and the near-Earth environment represent the region in which space weather plays out (courtesy of NASA).

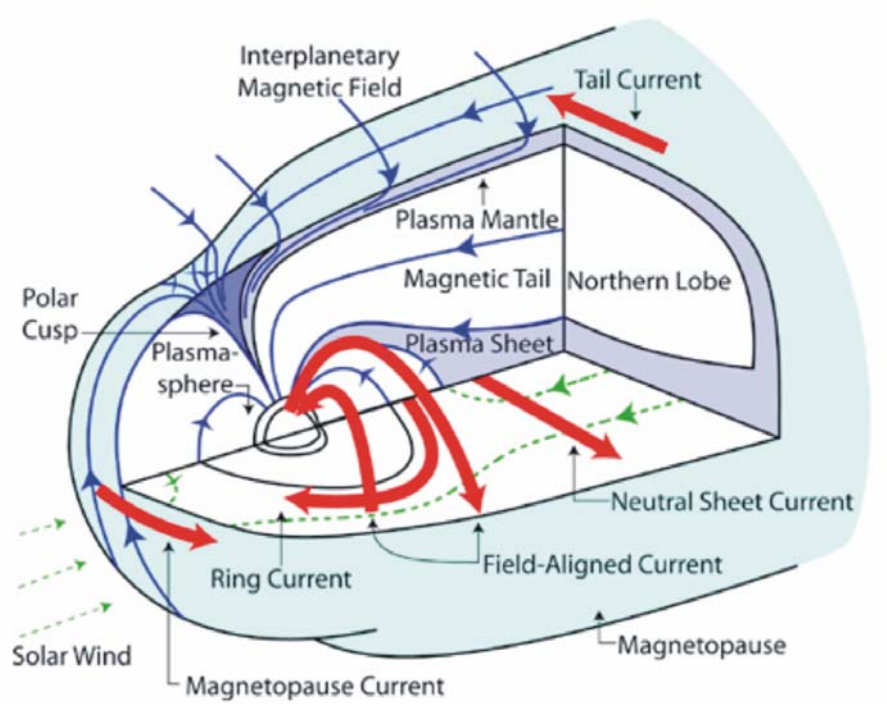


Fig. 3. The near-earth space environment showing the magnetosphere and many of the key plasma regions and current systems.

large-scale current systems and there are key regions of distinct plasmas (often separated – at least conceptually – by boundary layers). Trapped energetic particles constitute the van Allen radiation belts and a cold plasma region in the inner magnetosphere is called the plasmasphere. These plasma regions extend along magnetic field lines and couple into the ionosphere and the neutral atmosphere. Other chapters in this book will treat many aspects of the Sun-Earth system in great detail.

A point to bear in mind from Figs. 2 and 3 is that virtually all human technological systems operate on (or near) the Earth's surface or else in near-Earth space. Thus, power grids, communications systems, navigation satellites, and military space assets are all within – and are very much affected by – solar and magnetospheric disturbances. In this sense, space weather is an ever-present set of factors for advanced human technological resources. This chapter provides an overall introduction to space weather consequences and mitigation strategies.

2 Space Weather Effects

As shown in Fig. 4, the Sun can emit giant clouds of ionized gas (coronal mass ejections, CMEs) which contain upwards of 10^{16} grams of hot plasma. These CMEs can move outward from the Sun's surface at speeds of 1000 km/s (or more) and can have embedded within them strong magnetic fields and highly energetic particle fluxes. The active Sun is also the source of powerful solar flares and streams of high-speed solar wind flows. As these solar disturbances reach the Earth and its vicinity, they can give rise to long-lasting and disruptive disturbances called geomagnetic storms. High-energy ions and electrons produced during geomagnetic storms, as well as fluctuating magnetic fields themselves, can have detrimental effects on Earth-orbiting spacecraft and on humans in space (Lanzerotti, 2001 [12]).

As shown in Fig. 5a, high-energy protons and heavier ions arriving in near-Earth space can interact with spacecraft in several damaging ways. The ionization track that energetic ions can leave in microminiaturized electronics can upset spacecraft computer memories and can otherwise disrupt sensitive space electronics. The result can be damage to satellite solar power panels, confusion to optical tracker systems, and scrambling of spacecraft command and control software. Even more worrisome is the fact that high-energy solar particles can be damaging, or even potentially deadly, to astronauts who are in space at the time of major solar particle events (Turner, 2000 [19]).

Another aspect of the space environment that can be quite harmful to spacecraft is very energetic ("relativistic") electrons. As shown in Fig. 5b, these energetic electrons can penetrate through even thick spacecraft shielding and can bury themselves within dielectric (insulating) materials deep within spacecraft systems and subsystems. When sufficient charge has built up within dielectric materials such as coaxial cables or electronics boards, a

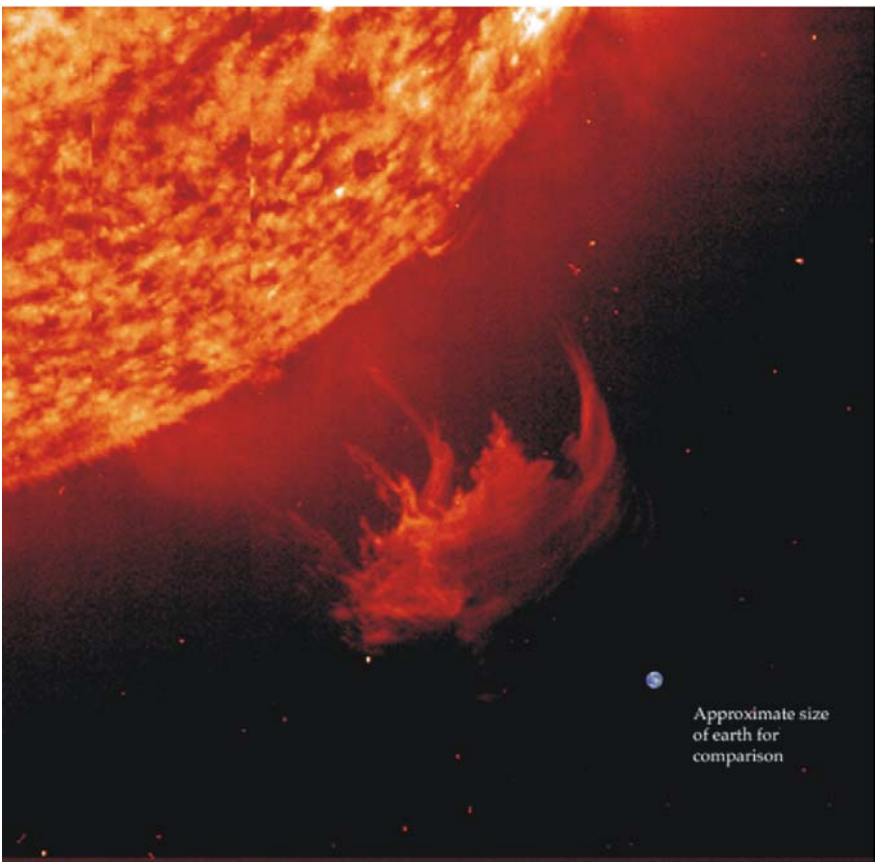


Fig. 4. A diagram illustrating a coronal mass ejection and also showing the Earth on a relative scale (courtesy of NASA).

powerful internal electrical discharge can occur (Baker, 1998 [2]; Robinson, 1989 [15]). This is very much like a miniature lightning strike within sensitive spacecraft electronics. Numerous recent spacecraft failures have been laid at the feet of this "deep dielectric charging" mechanism (Vampola, 1987 [20]; Baker, 1998 [2]; Baker et al., 1998 [8]).

Yet another space weather phenomenon of concern, known as "surface charging", is illustrated by part (c) of Fig. 5. Electrical charges coming from 10-100 kilovolt electrons within Earth's magnetosphere can accumulate on insulating surfaces of satellites. As with interior spacecraft insulators, if enough charge builds up on a region of surface dielectric material there can be a powerful, disruptive discharge. This can generate electrical signals in the spacecraft vicinity that can scramble and disorient the satellite and its subsystems (Robinson, 1989 [15]).

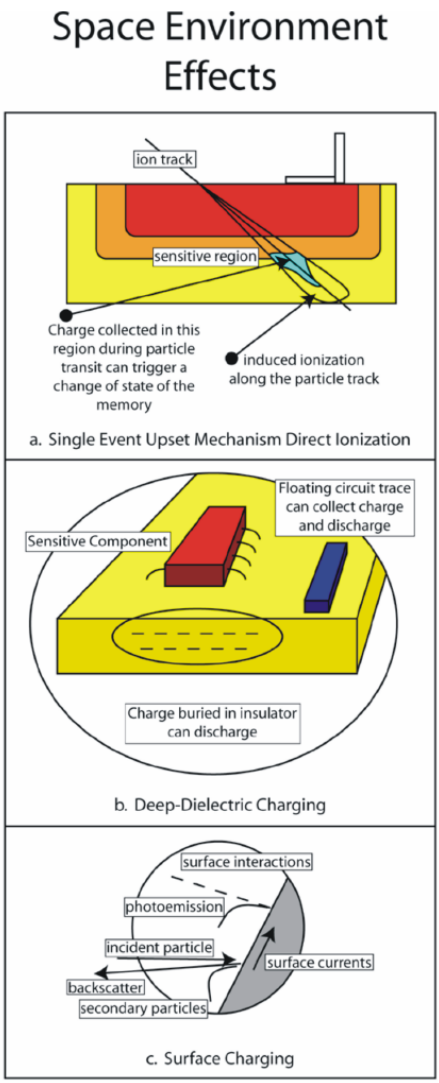


Fig. 5. A diagram illustrating space environment effects due to (a) Ions causing single-event upsets, (b) deep-dielectric charging, and (c) surface charging (adapted from Baker, 1998 [2]).

It is becoming increasingly understood and appreciated that continental-scale power generation and distribution systems are also vulnerable to the effects of space weather (Kappenman, 2001 [11]). Space storms can impact the operational reliability of electric power systems. For example, a major storm in 1989 shut down the Hydro Quebec power system in Canada for many hours. Space storms can disrupt power grids by introducing geomagnetically-

induced currents (GICs) into the transmission network. The GICs which flow through transformers, power lines, and grounding points can sometimes disrupt large portions of the power distribution system and such disruptions can occur within remarkably short periods of time (Kappenman, 2001 [11]). There are many other effects of space weather that manifest themselves in both subtle and very obvious fashions. A major space storm can modify the ionosphere of the Earth and therefore change the wavelength at which high-frequency (HF) radio communication is possible. This is a problem to the military and to airlines that are attempting to communicate with aircraft flying transpolar routes. Space weather can also cause sudden, unexpected heating of the Earth's upper neutral atmosphere. This heating causes an expansion of the upper atmospheric layer (the thermosphere) which can suddenly increase the drag force on low-altitude spacecraft (Lanzerotti, 2001 [12]; Singer et al., 2001 [18]).

3 Energetic Electrons and Space Weather

As illustrated in Fig. 5b, very high-energy electrons can penetrate through spacecraft walls and through electronics boxes to bury themselves in various dielectric materials (e.g. Robinson, 1989 [15]). This can, in turn, lead to electric potential differences in the region of the buried charge. In some instances, intense voltage breakdowns can occur leading to surges of electrical energy deep inside circuits. This can cause severe damage to various subsystems of the spacecraft.

Many examples of such "deep-dielectric charging" have been presented by various authors (e.g., Vampola, 1987 [20]; Baker et al., 1987 [7]). An interesting case study presented by Baker et al. (1987) [7] is shown in Fig. 6. In this figure, smoothed daily averages of $E = 1.4 - 2.0$ MeV electron fluxes at geostationary orbit are plotted versus time (late 1980 through early 1982). Also shown by bold vertical arrows are some of the main occurrences of star tracker anomalies onboard this geostationary operational spacecraft. The star tracker upsets were normally associated with high intensities of relativistic electrons. However, some high intensity electron events did not produce star tracker anomalies (see Baker et al., 1987 [7]) so there are more subtle controlling factors as well. Figure 7 shows how electrons must build up in dielectric materials for quite some time before a harmful discharge can occur. Thus, it is both the intensity of relativistic electron irradiation and its duration that is important. During some intense events in late 1981, the star trackers were actually turned off and so no operational "anomalies" could be recorded. The anomalies tended quite clearly to occur only during relatively long-duration events. Thus, it was not only the peak intensity of electrons, but also the duration of exposure that proved to be important.

Numerous previous studies (e.g., Reagan et al., 1983 [14]; Robinson, 1989 [15]; Wrenn, 1995 [22] have shown the clear role-played by high-energy elec-

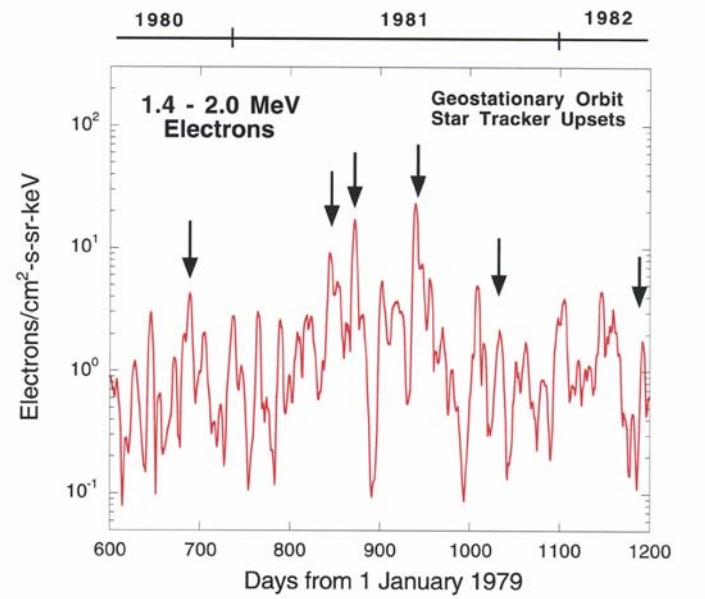


Fig. 6. Fluxes of 1.4-2.0 MeV electrons at geostationary orbit from late 1980 through early 1982. High electron flux events tended to be associated with star tracker anomalies (vertical arrows) on the spacecraft (from Baker, 2001 [4])

.

trons in many classes of spacecraft operational problems. Moreover, the quantitative level of radiation needed to produce deep-dielectric discharges has been rather clearly established in laboratory and spacecraft studies (e.g., Vampola, 1987 [20]). Figure 8, for example, adapted from Vampola's work shows results from the SCATHA mission that operated near geostationary orbit during the late 1970s and early 1980s. Deep dielectric discharges were monitored onboard the spacecraft and the daily fluences of E >300 keV electrons were concurrently measured. The probabilities of discharges went up dramatically when daily fluences exceeded 10^{11} electrons/cm^2. Above 10^{12} electrons/cm^2, the probability of discharges approached unity.

4 Magnetospheric Substorms

A significant effect of moderate geomagnetic activity ("magnetospheric substorms") from the standpoint of space operations is the occurrence of spacecraft surface charging (see Rosen, 1976 [17]. During a surface-charging event, insulated regions on a spacecraft may charge to several kilovolts potential (usually negative relative to the ambient potential). This charging occurs because of a lack of current balance between the ambient plasma medium and

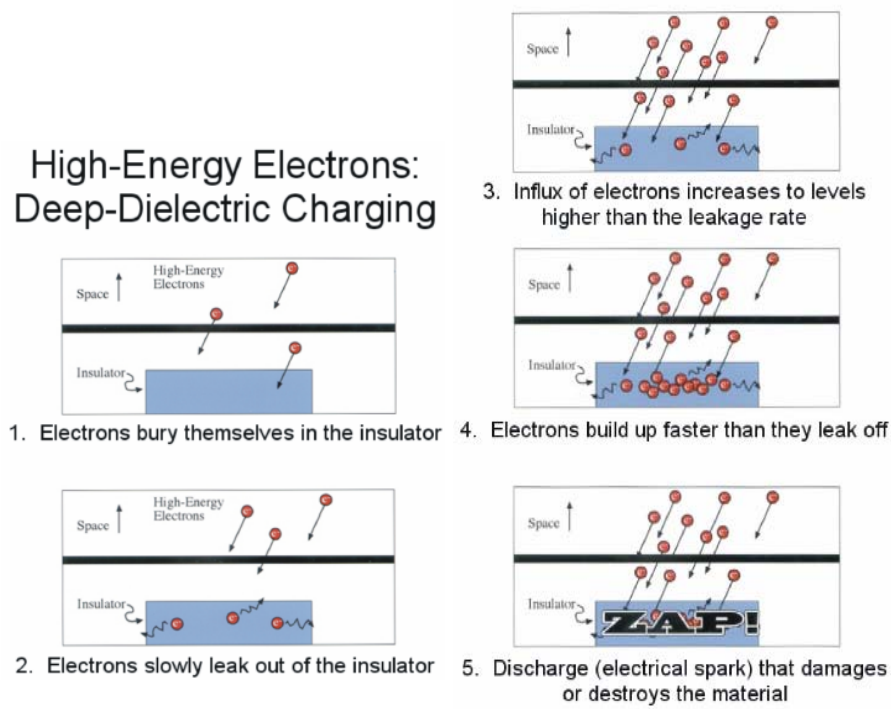


Fig. 7. A sequence of illustrations that show how high energy electrons can penetrate into buried material and cause deep-dielectric charging. The ultimate discharge (panel 5) can be very damaging to spacecraft subsystems (courtesy of G.D. Reeves).

the spacecraft surface (as illustrated in Fig. 5c). When a spacecraft is immersed in a cool, dense plasma, the incident particles (electrons and ions), as well as secondary emitted particles, photoelectrons, and backscattered electrons, all balance. This gives a low net spacecraft potential. However, in a very hot, tenuous plasma, current balance can be difficult to achieve and large potentials can build up.

Figure 5c shows the interaction at the surface of a spacecraft. This points out that there are currents near the surface of the spacecraft due to incident, backscattered, and photo-emitted particles. These various populations can, in principle, be examined to calculate the charge configurations for a given spacecraft. A sheath region that forms around the spacecraft is a volume strongly affected by the spacecraft. The plasma there is distorted by electric fields due to the charge of the spacecraft. The sheath region can also be affected by activity on the spacecraft such as thruster firings which extend the influence of the spacecraft farther into the plasma (e.g., Robinson, 1989 [15]). The sheath is complex in shape and depends on the motion of the spacecraft

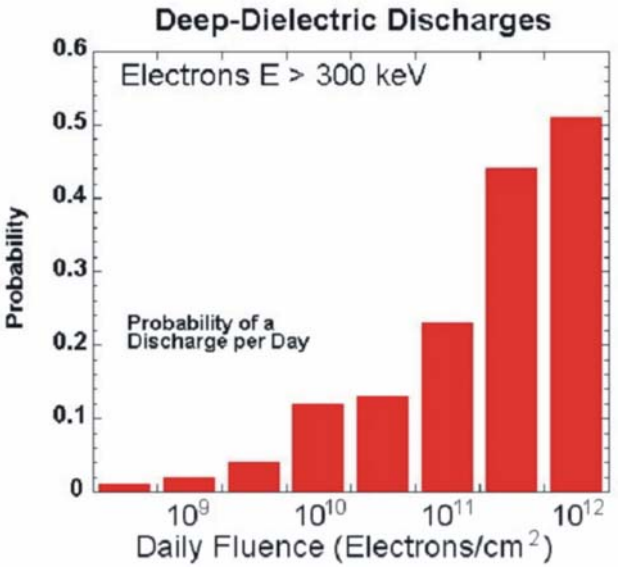


Fig. 8. Experimental results from Vampola (1987) [20] showing the probability of observing a dielectric discharge event as a function of the daily-integrated flux(fluence) of electrons with energy $E > 300$ keV (from Baker, 2000 [3]).

through the plasma as well as the plasma properties and the surface materials of the spacecraft. From an operational standpoint, differential charging of spacecraft surfaces that can lead to discharges. Discharges introduce noise into the system and may interrupt normal spacecraft operation, or represent a false command. In the process of discharge breakdown, physical damage may occur. This, in turn, may change the physical characteristics (thermal properties, conductivity, optical parameters, chemical properties, etc.) of the satellite. Furthermore, the release of material from the discharge site has been suggested as a contamination source for the remainder of the spacecraft (see Baker, 1998 [2] and references therein).

Figure 9, adapted from data presented in Rosen (1976) [17], shows the number of spacecraft anomalies detected at geostationary orbit as a function of spacecraft local time (LT). The anomalies include logic upsets as well as other significant operational problems for both military (Defense Support Program, DSP, and Defense Satellite Communications System, DSCS) and commercial (Intelsat) spacecraft. As may be seen, there is a very strong local time asymmetry in the number of anomalies with the vast majority occurring roughly between local midnight and local dawn. This is where substorm-injected electrons are seen most prominently (e.g., Baker (1998) [2], and references therein) and the LT distribution shown in Fig. 9 supports the view that surface charging has constituted a major cause of operational anomalies near geostationary orbit.

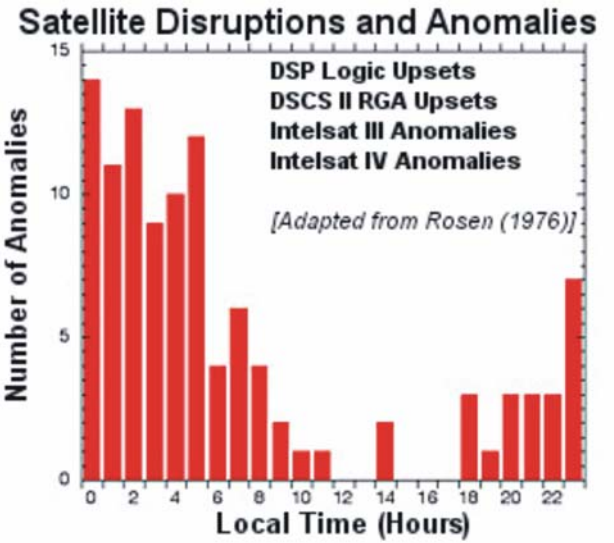


Fig. 9. Local time (LT) distribution of satellite disruptions and anomalies showing a strong occurrence frequency peak in the midnight and local morning hours (data from Rosen, 1976 [17]).

5 Space Weather Effects on the Electric Power System

Abelson (1996) [1] discussed the extensive changes that have recently characterized the electric power industry. He pointed out that modern, sophisticated end-users of electric power are growing increasingly susceptible to even momentary fluctuations of power. The standards for quality of the power supply continue to increase with the increased sophistication of society's utilization of electricity. He noted that interruptions and voltage sags cost users some $3 to $5 billion dollars per year and he went on to discuss new methods and technologies that may help to detect and correct for impending power quality problems.

The ever-evolving power grid is a complex and unique network of systems in which the production and delivery of the electrical energy all occurs at the speed of light. The design of the grid always took into consideration an array of contingency events and environmental challenges such as severe lightning, wind, and winter storms. Moreover, power grids have developed into continent-wide networks in which large blocks of power can be economically brokered over long distances and operation can be controlled by high speed devices. Notably, however, the size and complexity of the modern grid has introduced new vulnerabilities from the Sun (Kappenman, 2001 [11]).

Solar activity and eruptions commonly associated with the sunspot cycle can produce magnetic storms on Earth which have proven to have increasing impacts on technology systems such as electric grids as these systems be-

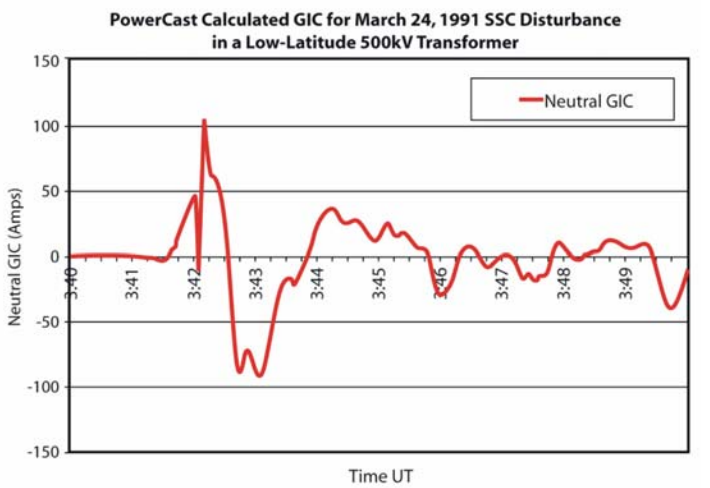


Fig. 10. PowerCast™ Models of a power grid can calculate GIC flows in every transformer for actual and hyopothetical storm scenarios. This is an example from a large SSC in March 1991 (courtesy of J. Kappenman, Metatech Corp.).

come more sophisticated. The example noted previously was a March 1989 geomagnetic storm that resulted in a day-long outage to the entire province of Quebec. This occurred due to a complex chain of events stemming from the unanticipated interaction of thyristor switched voltage regulators (SVC's) which culminated only one and a half minutes later in a complete blackout across the province. Further, this storm had the potential to drive a blackout that could have covered an area extending from Washington DC up through the New England states, according to an assessment made by the North American Electric Reliability Council. Figure 10 shows a large current surge (GIC) associated with a similarly powerful storm in March of 1991.

An Oak Ridge National Laboratory study indicated that a disturbance of this scale could have resulted in a potential economic cost of $3-6 billion. In similar effects on other technology systems (defense, communication and satellite systems) equally substantial costs and compelling impacts can be felt. Since these extreme storms can simultaneously affect entire continents, their impacts are in a category equivalent to large hurricanes and/or the San Francisco earthquake as it effects the reliability of the power grid (J. Kappenman, private communication).

Given the importance of reliable and uninterrupted power throughout the world, it is concluded that more effort needs to be made to provide warnings of geomagnetic disturbances (Baker and Kappenman, 1996 [6]). Reliable advance warnings would allow those impacted by storms the ability to prepare for these disturbances. In fact, this is the consensus of a joint commerce, research, and military task force under the sponsorship of the National Science

Foundation that was convened in 1994 to formulate a U.S. "National Space Weather Program Strategic Plan". This plan was issued in August 1995 and as noted in the executive summary, "The Nation's reliance on technological systems is growing exponentially, and many of these systems are susceptible to failure or unreliable performance because of extreme space weather conditions. We now have the scientific knowledge and the technical skills to move forward to dramatically improve space weather understanding, forecasts, and services to meet customer needs." The U.S. is now undertaking a multi-agency Space Weather Initiative (National Space Weather Program Strategic Plan, 1995 [13]) to provide early warning of impending space disturbances. This has the goal of devising methods to avoid power system failures and other impacts of the space environment on human technological systems. The space research and applications community has urged NASA, NOAA, NSF, and other government agencies to cooperate in the development of warning and amelioration methods to avoid catastrophic failures in power systems. Such warning methods could save millions or even billions of dollars each year. Figure 11 shows an example of the present state-of-the-art in specifying auroral current and GICs (courtesy of J. Kappenman).

6 Future Directions in Space Weather Studies

Space weather is, of course, not a new phenomenon. In a fascinating account written in the New Yorker magazine in 1959 (Brooks, 1959 [9]), John Brooks wrote about "The Subtle Storm". This was a major solar flare that commenced on February 9, 1958 and caused numerous problems with communications and other systems. Compared to 1958, the present-day world is immensely more complex and interconnected. The Earth's surface is crisscrossed by communication links, power grids, and a host of technological systems that did not even exist in 1958. When one considers the range of satellites orbiting the Earth from low to high altitudes, it is obvious that there is a complex "cyberelectric" cocoon that envelopes our entire planet and most elements of this web are susceptible in one way or another to space weather effects (Baker, 2002 [5]). Certainly, modern communication systems rely heavily on elements that include both ground links and satellite links. It is clear that world-wide communication systems can be detrimentally affected by adverse space weather (Lanzerotti, 2001 [12]; Baker et al., 1998 [8]; Singer et al., 2001 [18]. The failure and loss of even one key communication satellite – as occurred on May 19-1998 with the Galaxy IV satellite failure (see Fig. 12) – can affect millions or tens of millions of customers relying on telephones, pagers, and other communications technologies (Baker et al., 1998 [8]).

The first line of defense for human technology against the effects of space weather is, of course, to build robust systems that confidently withstand any space weather effects. To a large extent, this has already been done: Were it

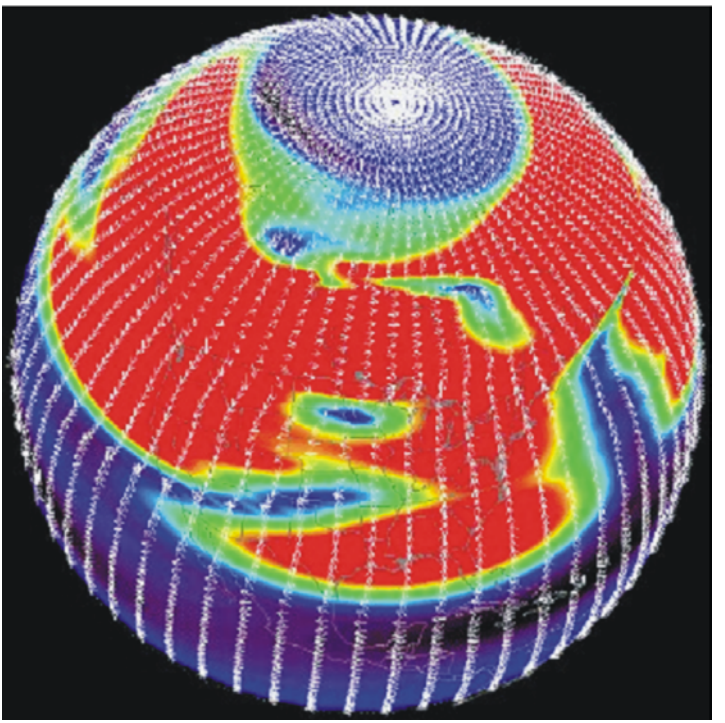
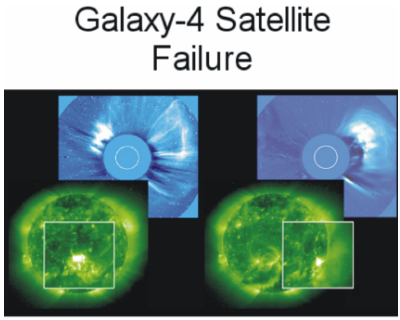


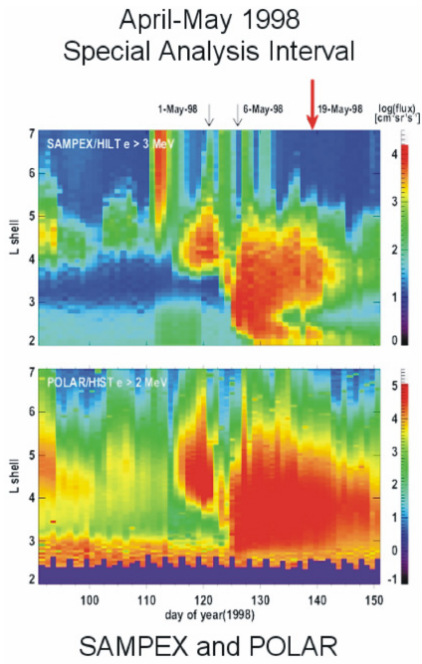
Fig. 11. A map showing magnetic field disturbances near the Earth's surface during a large geomagnetic storm on 15 July 2000 (courtesy of J. Kappenman, Metatech Corp.).

not so, there would be many more space weather-induced failures than we presently see. It is obvious that ground communication links, national power grids, and military installations – which must all withstand hurricanes, earthquakes, and floods – are very resilient and robust systems. Also, it is obvious that there are today many hundreds of satellites in Earth orbit fulfilling a wide variety of military and civilian purposes. Few of these fail catastrophically due to space weather. On the other hand, some spacecraft do fail suddenly due to space weather effects and nearly all spacecraft eventually fail due to the rigors of the hostile space environment. Thus, we need to know more about the nature of space weather elements, we need to specify better what the space environment is at any point in space, and we ultimately want to be able to predict (i.e., forecast) what the space weather environment will be anywhere in Earth's neighborhood many hours or days in the future. This is the goal of the U.S. National Space Weather Program (NSWP) (National Space Weather Program Strategic Plan, 1995 [13]; Robinsn and Behnke, 2001 [16]).

Space weather has become a major unifying theme and a uniting force for the entire solar-terrestrial research community. Understanding and predict-



a



b

Fig. 12. (a) Active regions on the Sun during April-May 1998, (b) and (c) The response of the Earth's radiation belt electrons to solar wind drivers. The time of the Galaxy-4 satellite failure is shown by the large vertical arrow (adapted from Baker et al., 1998 [8]).

ing such events is a challenge of great scope and complexity (Singer et al., 2001 [18]). The National Aeronautics and Space Administration (NASA) has now undertaken a major new initiative called "Living With a Star" (LWS) to observe systematically the disturbances arising on the Sun and to follow these space weather drivers all the way to their ultimate dissipation in Earth's atmosphere (Withbroe, 2001 [21]). The National Science Foundation (NSF) has also been a leading agency in the development of the National Space Weather Program (National Space Weather Program Strategic Plan, 1995 [13]; Robinsn and Behnke, 2001 [16]).The NSF has now selected a consortium of universities, industry partners, and national laboratories to form a Science and Technology Center dedicated to space weather. This "Center for Integrated Space-Weather Modeling" (CISM) is funded at several million dollars per year for the next 5-10 years and will have as its goal the building of physics-based models all the way from the Sun to the Earth's atmosphere. It involves numerous institutions (see Fig. 13) all around the U.S., and it works closely with the National Oceanic and Atmospheric Administration (NOAA).

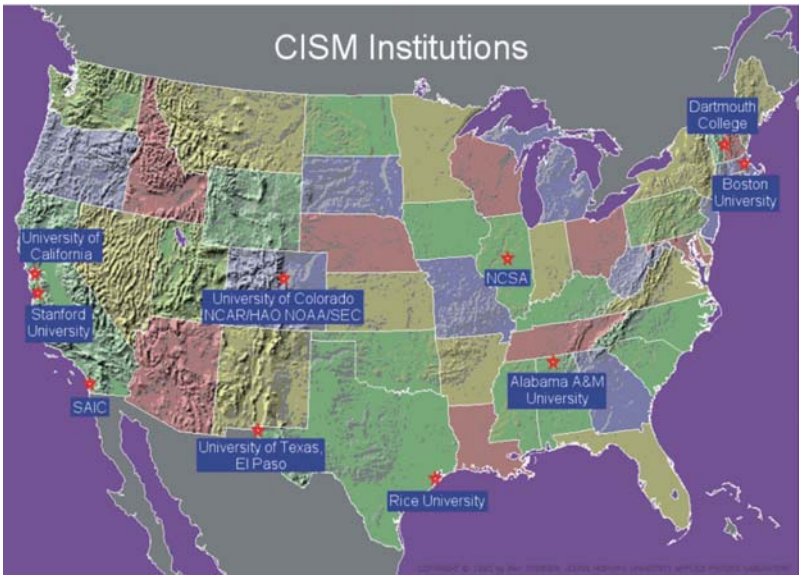


Fig. 13. Map of the U.S. showing the principal institutions involved in the NSF Center for Integrated Space Weather Modeling (CISM) (courtesy of W.J. Hughes).

7 Summary

Storms from the Sun are fascinating examples of energy transport and dissipation processes that are of undoubted importance in many cosmic contexts. Such storms are also beautiful when one observes the roiling surface of the Sun in soft X-rays or one stands outside on a dark night to observe the aurora borealis at northern latitudes. There is a splendor that accompanies space weather and there is also a danger from these powerful events that attracts and inspires popular readers (e.g., Carlowicz and Lopez, 2002 [10]). It is exciting to have the observational and modeling tools before us to be able to understand both the beauty and the threats presented by space weather. It is hoped that researchers from throughout the world can become engaged in space weather research. As shown by Fig. 14, agencies world-wide have now undertaken major space weather initiatives.

It is reasonable to expect that space weather research will continue and will, in fact, intensify over the next decades. Certainly, space-based and Earth-based human technology will remain susceptible to space weather effects. The increasing complexity and capability of human technology systems suggests that space weather will become more important as time goes on.

This chapter has given a very broad overview of several facets of space weather events and mechanisms. Subsequent chapters in this book go into much greater detail concerning the space environment, space weather mecha-

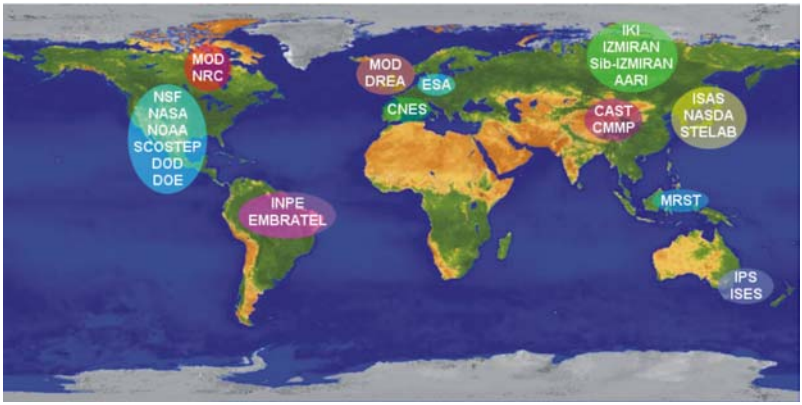


Fig. 14. A world-wide map showing agencies that are active in various aspects of space weather monitoring or forecasting (courtesy of J.H. Allen).

nisms, and technological consequences. The reader is urged to consult specific chapters if interested in a particular facet of space weather.

Acknowledgments

This research was supported by the Space Weather program of the U.S. National Science Foundation and by the Center for Integrated Space Weather Modeling (CISM). Portions of the work were also supported by NASA's SAMPEX and POLAR programs.

References

1. Abelson, P.H., Science, 271, 273, 1996.
2. Baker, D.N., Adv. Space Res., 23, 1,7,1998.
3. Baker, D.N., IEEE Trans. on Plasma Science, 28, 6, 2000.
4. Baker, D.N., Satellite anomalies due to space storms, in: Space Storms and Space Weather Hazards, I.A. Daglis, editor, Kluwer Academic Publishers, 2001.
5. Baker, D.N., Science, 297, 2002.
6. Baker, D.N., and J.G. Kappenman, Science, 273, p. 168, 1996.
7. Baker, D.N., et al., J. Electrostatics, 20, 3, 1987.
8. Baker, D.N., J.H. Allen, S.G. Kanekal, and G.D. Reeves, EOS, 79, 477, 1998.
9. Brooks, J., "Reporter at Large: The Subtle Storm", The New Yorker, 27 Feb. Issue, p. 39, 1959.
10. Carlowicz, M.J., and R.E. Lopez, Storms from the Sun, Joseph Henry Press, Washington, DC, 2002.

11. Kappenman, J.G., in: Space Storms and Space Weather Hazards, Chap. 13, p. 335, Kluwer Acad. Pub., 2001.
12. Lanzerotti, L.J., in Space Weather, Geophys. Monograph 125, p. 11, Amer. Geophys. Union, Washington, DC, 2001.
13. National Space Weather Program Strategic Plan, Office Fed. Coord. For Met. Services, NOAA, Silver Spring, MD (1995).
14. Reagan, J.B., et al., ISEE Trans. Elec. Insul., E1-18, 354, 1983.
15. Robinson, P.A., Jr., Spacecraft environmental anomalies handbook, JPL Report GL-TR-89-0222, Pasadena, CA, 1989.
16. Robinson, R.M., and R.A. Behnke, in Space Weather, Geophys. Monograph 125, p. 1, Amer. Geophys. Union, Washington, DC, 2001.
17. Rosen, A. (editor), Spacecraft charging by magnetospheric plasmas, AIAA, 47, New York, 1976.
18. Singer, H.J., et al., in Space Weather, Geophys. Monograph 125, p. 23, Amer. Geophys. Union, Washington, DC, 2001.
19. Turner, R., IEEE Trans. on Plasma Science, 28, 2103, 2000.
20. Vampola, A.L., J. Electrostat., 20, 21, 1987.
21. Withbroe, G.L., in Space Weather, Geophys. Monograph 125, p. 45, Amer. Geophys. Union, Washington, DC, 2001.
22. Wrenn, G.I., J. Spacecraft and Rockets, 32, 514, 1995.

The Sun and Its Restless Magnetic Field

Manfred Schüssler

Max-Planck-Institut für Sonnensystemforschung, Max-Planck-Str. 2,
37191 Katlenburg-Lindau, Germany

Abstract. The permanently changing magnetic field of the Sun is the ultimate driver of the various effects refered to as *space weather*. This chapter gives an overview of the physics of the solar magnetic field and its variability. This includes a brief description of the structure of the Sun, and a discussion of the origin of the solar magnetic field.

1 Introduction

The permanently changing magnetic field of the Sun is the ultimate driver of the various effects summarized under the labels *solar-terrestrial relations* and *space weather*. The aim of this chapter is to give a brief overview of the physics of the solar magnetic field and its variability. This area of research is vast, so it would be pointless to aim at a comprehensive and detailed account in a contribution of this format. So I restrict myself to a summary of the basic observational results and the currently prevailing theoretical models for the various processes. Reference is made to more detailed accounts whenever possible. More comprehensive representations of relevant topics can be found, for instance, in the books of Stix (2002) [1], Schrijver and Zwaan (2000) [2], Priest (1984) [3], Cox et al. (1991) [4], and Schmelz and Brown (1992) [5].

This contribution is organized as follows. After briefly presenting the general structure of the Sun in Sect. 2, various aspects of the solar magnetic variability will be discussed in Sect. 3: magneto-convection (3.1), active regions (3.2), the global transport of magnetic flux on the solar surface (3.3), the 11/22 year cycle and its long-term modulation (3.4). Section 4 is devoted to the origin of the magnetic field and the activity cycle. We conclude with a brief outlook in Sect. 5.

2 The Structure of the Sun

Schematically, the Sun and its extended atmosphere can be divided in a number of layers according to the physical processes dominating the respective regions.

Theoretical models of the solar interior, its pressure and temperature stratification as a result of hydrostatic equilibrium and the processes for

M. Schüssler, The Sun and Its Restless Magnetic Field, Lect. Notes Phys. **656**, 23–49 (2005)
http://www.springerlink.com/

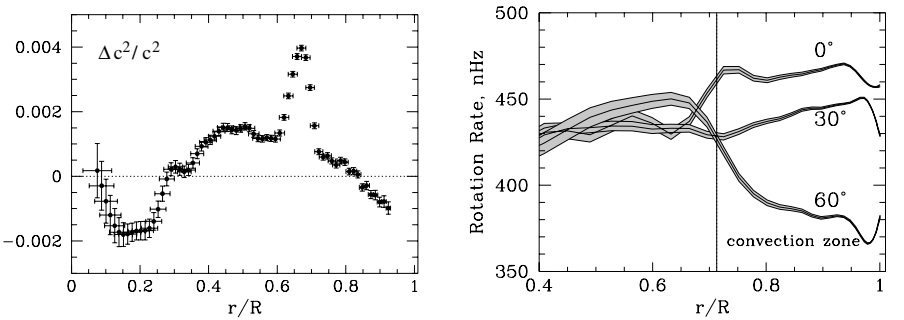


Fig. 1. Left: Relative difference between the (squared) sound speed profile in the solar interior as inferred from helioseismology and a theoretical standard solar model (Christensen et al., 1996 [7]). Right: Solar rotation rate inferred from helioseismology as a function of radius at three solar latitudes. The formal errors are indicated by the shaded regions (both figures from from Kosovichev et al. (1997) [8]).

the generation and transport of energy have reached a large degree of sophistication. The validity of these models can be empirically tested through *helioseismology*, the analysis of the acoustic eigenmodes of the solar body (Christensen and Dalsgaard, 2003 [6]). Similar to the information about the Earth's interior gained through the detection of seismic waves from earthquakes, the internal structure of the Sun can be reconstructed fairly precisely on the basis of the measured frequencies of the acoustic eigenmodes excited by solar convection. The solar structure as derived through helioseiosmology is in very good agreement (at a level of 0.1%) with the theoretical models (see Fig. 1, left panel). Locally confined larger deviations, for instance directly below the convection zone, indicate that more subtle additional processes like gravitational Helium settling have to be taken into account.

The core of the Sun (out to about 15% of the solar radius, $R_\odot$) is the domain where thermo-nuclear reactions operating at temperatures above 10^7 K generate the solar energy flux. Most important is the proton-proton (pp) cycle, which (stepwise) generates ^{4}He from hydrogen nuclei (protons). The neutrinos produced in the course of the nuclear reactions leave the core almost uninhibited and can be detected on Earth as direct witnesses of the energy-generating processes. The long-standing 'solar neutrino problem', resulting from the detection of much less (electron) neutrinos than expected from the nuclear reactions, found its resolution in the existence of neutrino oscillations which periodically change the 'flavor' of the neutrinos during their flight to Earth, so that only part of them are registered in the detctors (Bahcall et al., 2003 [9]). This implies also a finite (albeit small) rest mass of the neutrinos and thus entails modifications of the standard model of elementary particles. The good agreement between helioseismic results and theoretical models of the solar interior indicated early on that the solar neutrino problem most

probably would find its resolution from the side of particle physics, not solar physics.

The energy generated in the core is transported outward by photons in the radiation zone of the Sun (often also called radiative interior), which covers the region 0.15–0.7 $R_\odot$. The photons are so often absorbed and re-emitted in the dense solar matter that the transfer of energy through the radiation zone with its outward decreasing temperature can be very well be approximated by a diffusion process with a characteristic timescale of about 10^7 years.

As the temperature decreases outward more and more, electrons start to recombine with the various atomic nuclei. The resulting possibilities for electronic transitions increase the opacity of the matter and thus decrease the diffusion coefficient for radiation, so that the temperature gradient has to grow steeper in order to keep the energy flux constant in a steady state. Around 0.7 $R_\odot$ and at $T \simeq 2{\cdot}10^6$ K, the required temperature gradient for the radiative transport of the energy becomes steeper than the adiabatic temperature gradient. As a result, the stratification becomes unstable with respect to convective motions, which rapidly take over as the dominant mechanism of energy transport. The very good mixing properties of the convective motions lead to an almost adiabatic (isentropic) stratification of the convection zone, which spans the radius interval 0.7–1.0 $R_\odot$ (Schüssler, 1992 [10]).

Since convective motions can penetrate a certain depth into the stable radiation zone, a convective overshoot layer forms. The slightly subadiabatic (stable) stratification of the overshoot layer favors the storage of magnetic flux (Moet al., 1992 [11]). As we shall see in more detail Sect. 4.1, this is one of the reasons why the interface between the convection zone and the radiative core at about 0.7 $R_\odot$ is of particular interest for the generation of magnetic flux by the solar dynamo. A second indication comes from helioseismology, which can be used to infer the rotation rate in the solar interior as a function of depth and latitude. As shown on the right panel of Fig. 1, there is a layer of strong radial rotational shear (differential rotation) straddling the transition between the radiation zone and the convection zone (Schou et al., 1998 [12]). Since differential rotation is one of the basic ingredients of virtually all models for the solar dynamo mechanism (which maintains the solar magnetic field in the first place), the discovery of this shear layer (the so-called solar tachocline) has important consequences for our understanding of the solar dynamo. The other ingredient for the solar dynamo are the convective motions of the electrically conducting solar plasma (see Sect. 4).

The mass motions of solar convection can directly be observed in the form of granulation on the visible solar surface, which is the layer where the solar plasma becomes transparent for visible light. Figure 2 shows the granulation pattern between sunspots and pores, which are cooler and darker than their surroundings because their strong magnetic field suppresses the energy-carrying convective motions. The bright granules correspond to rising hot gas while the network of darker intergranular lanes represents the downflow regions where the matter descends again into the convection zone

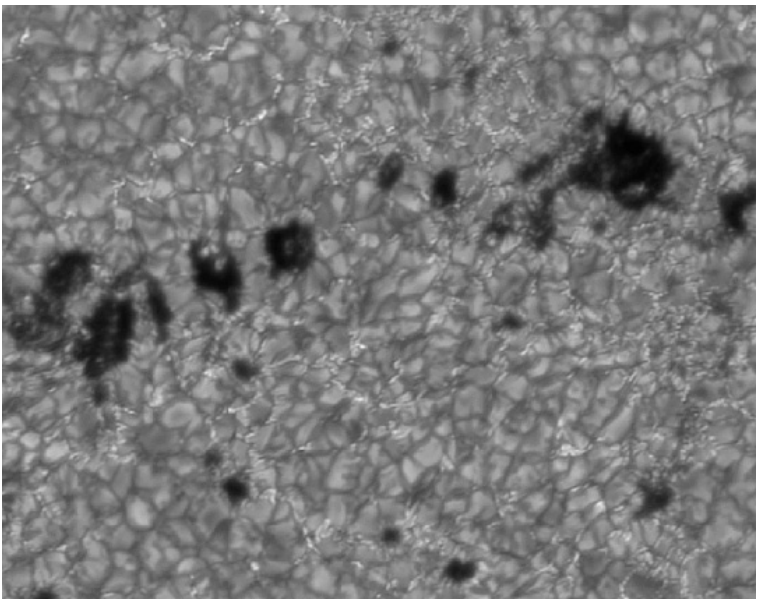


Fig. 2. Granulation, sunspots, pores, and magnetic elements in the solar photosphere as visible in the Fraunhofer G band at 430.4 nm, a wavelength range dominated by lines formed by the CH molecule. The magnetic flux outside the dark pores resides in tiny field elements located in the intergranular downflow regions. While larger magnetic structures are darker than the average photosphere, small-scale magnetic concentrations appear bright owing to the reduced abundance of CH molecules (and thus less absorption in the molecular lines) in their hot and tenuous atmospheres (image taken with the Dutch Open Telescope at the Roque de los Muchachos Observatory on La Palma, Spain; courtesy P. Sütterlin).

after it has lost a substantial part of its internal energy through radiation into space. Apart from the dark sunspots and the granulation pattern, Fig. 2 also shows conspicuous bright points in the intergranular lanes: these represent magnetic flux concentrations with field strengths similar to those in sunspots ($\simeq 1500\,\mathrm{Gauss} = 0.15\,\mathrm{Tesla}$, see (Solanki, 1993 [13])). Such magnetic elements owe their existence to the interplay between convection, radiation, and magnetic field commonly referred to as solar magneto-convection (cf. Sect. 3.1).

The kinetic energy density of convection, the thermal and the magnetic energy density are all of the same order of magnitude in the solar photosphere comprising the visible surface and the first few hundred km of the overlying solar atmosphere. Above the photosphere, the exponential decrease of gas density and pressure rapidly leads to a situation in which the magnetic field completely dominates the structure and dynamics. In the chromosphere and in the corona, the energy built up in the magnetic field by the interaction

with convective motions in the lower atmosphere is released by various (not very well understood) processes like, for instance, shock dissipation, current sheet dissipation, and reconnection of magnetic field lines (Narain and Ulmschneider, 1990 [14]; Narain and Ulmschneider, 1996 [15]). As a consequence, the gas is heated and the temperature increases outward until coronal temperatures of a few million K are reached. At the same time, magnetic energy is also released in various kinds of large impulsive phenomena like flares, surges, or coronal mass ejections. Such events are the cause of the multitude of near-Earth phenomena and effects that travel under the signature 'space weather'.

In the corona, it is important to distinguish between the regions of 'open' and 'closed' magnetic topology (see Fig. 3). In regions of open field lines, which extend into the far reaches of interplanetary space and heliosphere, much of the mechanical energy delivered from the lower atmosphere goes into the acceleration of a fast solar wind. Such open regions exhibit a low mass density, show weaker emission of radiation, and therefore appear relatively dark (coronal holes). On the other side, the magnetically closed regions are dominated by magnetic loops containing dense, hot, and bright plasma. The lower latitudes of the Sun are dominated by closed regions owing to the many bipolar active regions emerging there (see Sect. 3.2). In contrast, the polar regions of the Sun are typically covered by large, magnetically open coronal holes. The detailed distribution of open and closed regions depends strongly on the phase of the solar cycle.

3 Solar Magnetic Variability

The most important timescale of solar magnetic variability is the 11-year activity cycle (or 22-year magnetic cycle) of active regions and sunspots. The underlying large-scale magnetic flux systems are believed to be generated by a hydromagnetic dynamo mechanism based upon convective motions and (differential) rotation (see Sect. 4). In addition to the conspicuous 11/22-year cycle, the solar magnetic field is variable on all temporal and spatial scales on which it has been observed so far. The timescales range from minutes to centuries with different physical processes underlying these variations. Magneto-convection, i.e., the interaction of convective motions and magnetic field, is responsible for the variability on timescales between minutes and days. The evolution of active regions and global flux transport over the solar surface covers the timecales between days and several years, while the range from decades to centuries is governed by the dynamo process and its long-term modulation.

Besides the effects that are relevant for space weather, like flares and coronal mass ejections, the magnetic variability of the Sun causes variations of its UV and X-ray flux as well as in total irradiance and the flux of cosmic rays hitting the Earth's atmosphere, all of which my affect the terrestrial climate (Friis-Christensen et al., 2000 [16]; Wilson, 2000 [17]).

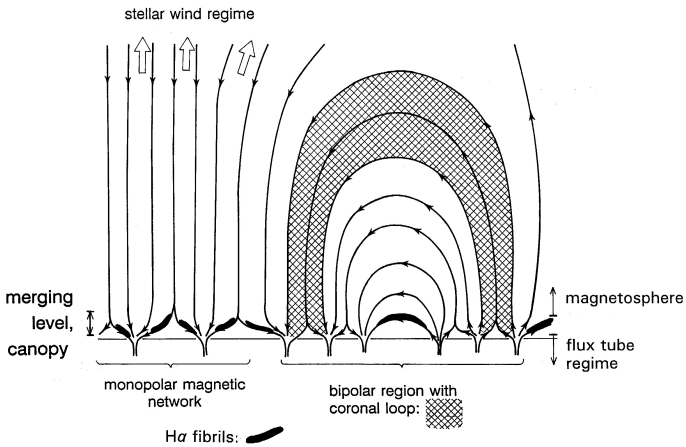


Fig. 3. Magnetic structure in the solar atmosphere. The intermittent magnetic flux concentrations in the photosphere expand upward and merge in the chromosphere. In the corona, the magnetic topology is either open (field lines extending into the heliosphere) or closed (field lines bending back to the photosphere, forming loop-like structures). Magnetically closed regions support larger plasma densities in coronal loops while open regions (coronal holes) permit the outflow of matter in the fast solar wind (from Schrijver and Zwaan (2000) [2]).

3.1 Magneto-convection

In an electrically well-conducting fluid (or, to be more precise, if the magnetic Reynolds number of the considered flow is very large compared to unity), we have the situation of magnetic flux freezing (Choudhuri, 1998 [18]). This means that plasma cannot change from one field line to another, so that either the field lines have to follow the motion of the plasma (weak field, no significant Lorentz force) or the magnetic field suppresses the motion perpendicular to the field lines (strong field, dominating Lorentz force).

The condition of flux freezing is clearly fullfilled for the convective flows observed near the solar surface. The instationarity of the convective flow patterns leads to a continuous rearrangement and reshuffling of magnetic flux, which moves along the lanes of the various downflow networks like a 'magneto-fluid'. Consequently, the length- and timescales of the various convective patterns are impressed upon the evolution of the magnetic field. These are granulation (which carries nearly all of the energy flux) with a timescale of about 5–10 minutes and a length scale of 1–2 Mm, mesogranulation with a timescale of a few hours and a length scale of about 5 Mm, and supergranulation with a timescale of 1–2 days and a length scale of about 20 Mm. The buoyant magnetic flux that crosses the solar surface is carried by the horizontal motions of these flow patterns and accumulates in the convective downflow regions, leading to a hierarchy of network patterns, the most prominent being the supergranular network. This process, known as flux expulsion

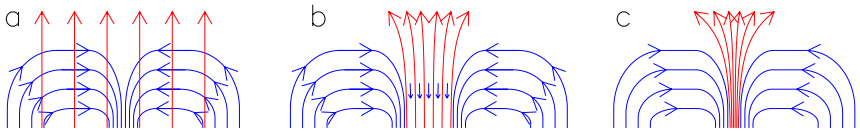


Fig. 4. Schematic sketch of the processes that lead to intensification of magnetic field in the solar (sub)photosphere. *a:* the horizontal flows of granular convection sweep magnetic flux towards the downflow regions (lighter vertical lines are magnetic field lines, the other set of lines represents stream lines of the flow). *b:* magnetic forces suppress the convective motions when the magnetic energy density becomes comparable to the kinetic energy density. This throttles the energy supply into the magnetic region, the gas continues to cool owing to radiative losses, and the internal downflow is enhanced. The superadiabatic stratification further amplifies the flow. *c:* the downflow has evacuated the upper layers and the magnetic field in the quenched tube has increased accordingly to maintain the balance between external and internal (gas plus magnetic) pressure.

(Proctor and Weiss, 1982 [19]; Hurlburt and Toomre, 1988 [20]), is capable of enhancing the field to about equipartition of the magnetic energy density with the kinetic energy density of the flow. For solar granulation, this corresponds to a field strength of about 500 G.

For the case of the solar (sub-)photosphere, however, the equipartition limit may be greatly exceeded through thermal effects. Since the horizontal flows of granular convection which sweep the magnetic field to the downflows also carry heat to those regions, the retardation of the flows by the growing magnetic field leads to a cooling of the magnetic region since the the radiative losses can no longer be balanced by the throttled horizontal flow. This drives an enhanced downflow of plasma along the field lines, which is further accelerated by the strongly superadiabatic stratification of the top layers of the convection zone. As a consequence, the upper layers of the magnetic structure become nearly evacuated and the surrounding plasma compresses the magnetic field until an equilibrium of total pressure (i.e., gas pressure plus magnetic pressure) is reached at field strengths in the kG range. In this way, the magnetic field can be locally intensified to values which are only limited by the confining pressure of the external gas. This convective intensification or convective collapse process (Parker, 1978 [21]; Spruit and Zweibel, 1979 [22]) is schematically sketched in Fig. 4. Observational evidence for the process has been reported (Zwaan et al., 1985 [23]; Bellot et al., 2001 [24]).

The collapse process loses its efficiency for flux tubes with a diameter smaller than the photon mean free path of about 100 km at the base of the photosphere (Venkatakrishnan, 1986 [25]). At that limit, the tube interior is kept at the temperature of its environment and the evacuating downflow is throttled, so that the equipartition limit cannot be exceeded. In fact, observations in the infrared spectral range reveal the existence of small magnetic

features with about equipartition field strength in the photosphere (Lin, 1995 [26]).

Sufficiently strong magnetic fields in large structures, like in sunspots or pores, suppress the convective motions and the associated energy transport, so that they become dark in comparison to the surrounding photosphere. The remaining energy flux (about 20% of the undisturbed flux) is still much too large to be carried by radiation alone. It is assumed that oscillatory convection in a strong magnetic field is the dominant process for energy transport in sunspot umbrae (Hurlburt et al., 1989 [27]; Weiss et al., 1996 [28]). Small-scale concentrations of strong magnetic field appear brighter than their surroundings because of lateral influx of radiation from the 'hot walls' around the low-density flux concentrations (Spruit et al., 1991 [29]) and, presumably, also by dissipation of mechanical energy in their higher atmospheric layers. Note that for a cylindrical flux tube the ratio of the heating surface area of its wall and the internal volume to be heated is inversely proportional to the tube radius. The hot walls of small flux tubes become best visible near the limb of the Sun, where the combined effect of many flux tubes gives rise to bright faculae (Topka et al., 1997 [30]). Averaged over the whole Sun, the enhanced brightness of the magnetic elements dominates over the reduced energy flux in the sunspots, so that the total solar radiative output increases with growing amount of magnetic flux in the photosphere (Fligge et al., 2000 [31]).

The observed general behaviour of magnetic flux in the solar atmosphere is reproduced in numerical simulations (Nordlund and Stein, 1990 [32]; Vögler et al., 2003 [33]; Vögler and Schüssler, 2003 [34]). Realistic simulations of solar magneto-convection require a detailed non-local and non-grey radiative transfer and need to include compressibility. In the solar photosphere and upper convection zone, the deviations from the ideal gas law due to partial ionization of the fluid are important since transport of latent heat contributes significantly to the total convective energy transport. The large range of spatial scales to be covered calls for simulation codes designed for use on parallel computers.

Figures 5 and 6 show some results from three-dimensional simulation runs with the MURAM (**M**ax-Planck-Institute for Aeronomy and **U**niversity of Chicago **Ra**diation **M**HD) code (Vögler and Schüssler, 2003 [34]). The dimensions of the computational domain are 1400 km in the vertical direction and 6000 km in both horizontal directions, with a resolution of $100 \times 288 \times 288$ grid points. The simulation starts from a plane-parallel model of the solar atmosphere extending from 800 km below to 600 km above the level of continuum optical depth unity at 500 nm. After convection (granulation) has fully developed, a homogeneous vertical initial magnetic field of 200 G (corresponding to the average field strength in asolar active region) is introduced. Within a few minutes of simulated time (approximately one turnover time of the convection) most of the magnetic flux is transported to the downflow

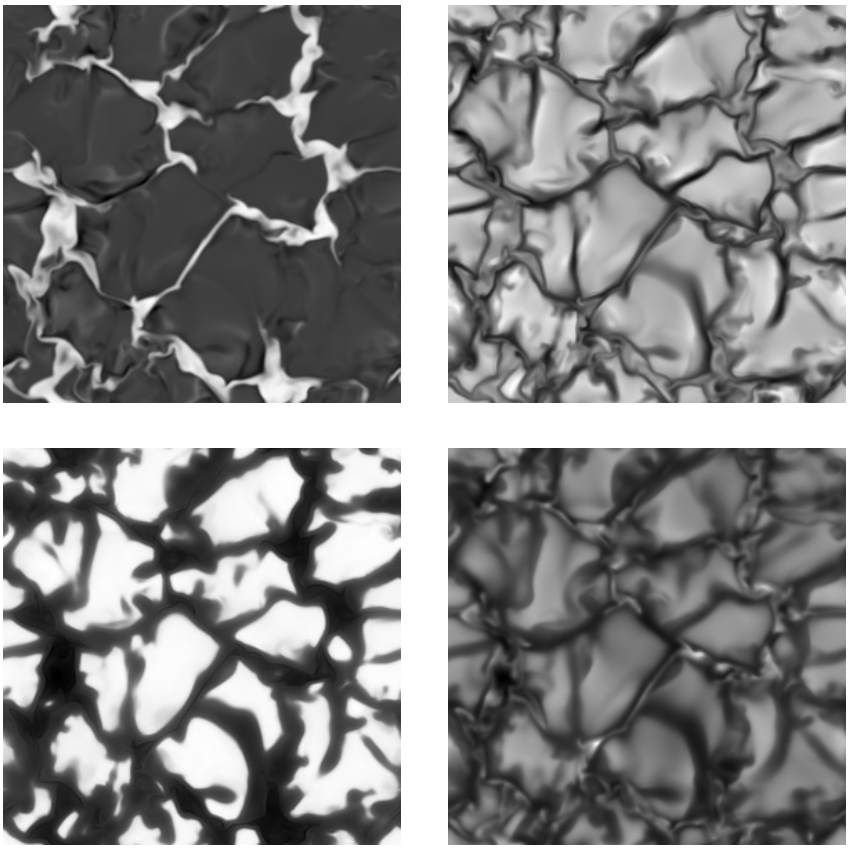


Fig. 5. Brightness map (lower right) and horizontal cuts at the average geometrical height corresponding to optical depth unity of vertical magnetic field (upper left), vertical velocity (upper right) and temperature (lower left). Light and dark shades indicate higher and lower values, respectively. The spatial dimensions correspond to 6000×6000 km on the Sun. The velocity plot shows granular upflows shaded in light grey separated by intergranular downflow lanes. In the magnetic-field plot, the strong sheet- and pore-like magnetic field concentrations appear in white. They are organized mainly in a 'meso-scale' network with a typical size significantly larger than the spatial scale of the granulation.

lanes of the granulation pattern. There the field is concentrated to kilogauss values by the convective intensification process described above.

For a snapshot taken about 2 hours solar time after the start of the magnetic phase, Fig. 5 shows the vertical magnetic field, vertical velocity, and temperature distributions on a horizontal plane corresponding roughly to the average level of optical depth unity at 500 nm (i.e., the visible solar sur-

face). In addition, the frequency-integrated intensity (brightness) is shown on the lower right panel. The magnetic map shows sheet-like magnetic structures extending along intergranular downflow lanes, while larger structures with diameters of up to 1000 km (comparable to small pores on the sun) are formed at granule vertices where several downflow lanes merge. Typical field strengths in these field concentrations at a height corresponding to optical depth unity are between 1500 and 2000 G.

The network of magnetic structures is organized on a 'mesoscale' which typically comprises 4–6 granules. While this magnetic pattern is rather stable (it evolves on a time scale much larger than the average lifetime of granules), the small-scale pattern of the field concentrations is highly time-dependent, with magnetic flux being constantly redistributed within the magnetic network. The appearance of a mesoscale is a typical and robust feature of convection simulations (Cattaneo et al., 2001 [35]).

In the brightness map shown in Fig. 5, the pore-like structures appear dark owing to the reduced efficiency of convective energy transport and hence lower temperature at optical depth unity. There is a considerable small-scale variation of brightness over their surface which is related to localized hot upflows in regions of reduced field strength. In the thin sheets, lateral heating effects in combination with the depression of the level of optical depth unity lead to a net brightening with respect to the surrounding downflow regions (see Fig. 6).

3.2 Active Regions

The magnetic flux related to the 11/22-year solar cycle emerges at the photosphere in the form of magnetically bipolar regions (Zwaan and Harvey, 1994 [36]) with a continuous size spectrum, which are best visible in magnetograms, maps of the photospheric magnetic field (see Fig. 7). The larger of these are the active regions, which usually contain sunspots and are restricted to about ±30 degree heliographic latitude. New active regions show a tendency to appear in the vicinity of previously emerged active regions, thus often forming sunspots nests and complexes of activity (Gaizauskas et al., 1983 [37]; Schrijver and Zwaan, 2000 [2]). The smaller ephemeral regions do not form sunspots and have a broader distribution in latitude (Harvey and Martin, 1973 [38]; Harvey, 1992a [39]; Harvey, 1993 [40]). While active regions can have lifetimes of up to several months, ephemeral regions typically have a very short decay time of the order of days. On the other hand, the rate of flux emergence in ephemeral regions is about two orders of magnitude larger than the corresponding rate from active regions (Schrijver et al., 1997 [41]; Hagenaar, 2001 [42]).

The activity cycle of ephemeral regions is more extended than that of active regions with sunspots and runs ahead in phase (Harvey, 1993 [40]; Harvey, 1994a [43]): ephemeral regions belonging to a given cycle (as distinguished, e.g., by emergence latitude) start emerging 2–3 years before the first

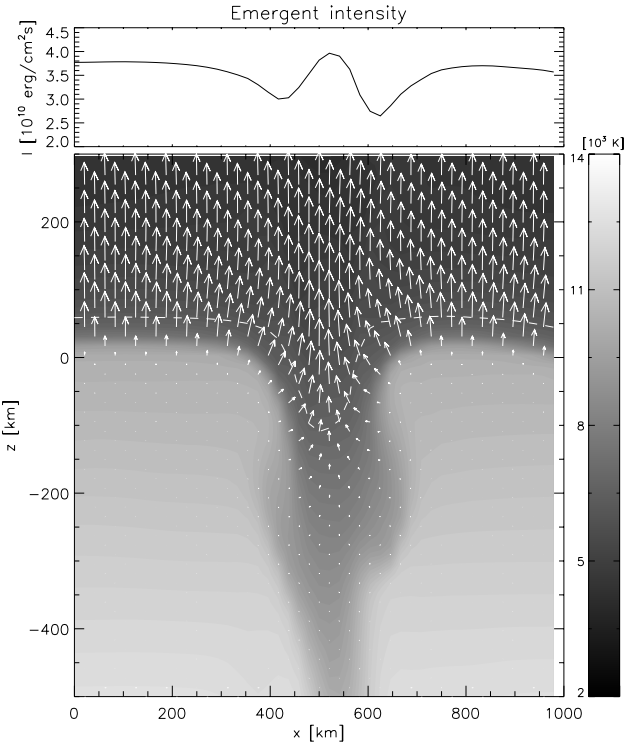


Fig. 6. Radiative properties of a simulated sheet-like magnetic structure in an intergranular lane. The lower part of the figure shows a grey-shading of the the temperature distribution in a vertical cut containing the flux sheet, together with the radiative flux vectors. There is an influx of radiation from the hot walls of the flux sheet into its partially evacuated interior. The radiative heating leads to enhanced brightness (shown in the upper part of the figure), so that the flux sheet appears as a bright structure within the darker integranular lane.

sunspots of the corresponding cycle appear. Therefore, we have an overlap of magnetic flux emergence in ephemeral regions from the old and from the new cycle around each sunspot minimum. This effect can lead to a secular variation of the total (cycle-related) magnetic flux at the solar surface (Solanki et al., 2002 [44]).

How do active regions originate? Many observed properties of sunspot groups (rapid formation of a well-defined bipolar structure, east-west orientation, Hale's polarity rules, Joy's law for the tilt angle of sunspot groups with respeect to the east-west direction) indicate that bipolar magnetic regions originate from a toroidal magnetic flux system, which is generated, amplified and stored in the vicinity of the tachocline near the bottom of the convection zone (Schüssler et al., 1994 [45]). Flux expulsion by convective flows (Proctor and Weiss, 1982 [19]) and the magnetic Rayleigh-Taylor

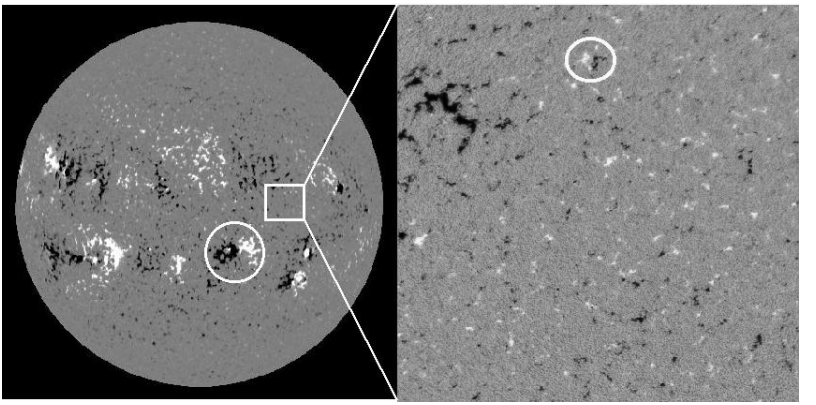


Fig. 7. Map of the magnetic field in the solar photosphere taken with the Michelson Doppler Interferometer (MDI) onboard SOHO. Black and white indicate positive and negative magnetic polarity, respectively, grey represents low magnetic flux. The dominant features are bipolar active regions (an example is circled in the left-hand panel) and extended unipolar domains with a network structure outlining the convective flow pattern of supergranulation. The panel on the right-hand side shows (schematically) a blow-up with small-scale flux of mixed polarity and an ephemeral region (circled).

instability (Cattaneo and Hughes, 1988 [46]; Fan, 2001 [47]) lead to the formation of intermittent magnetic structure in the form of magnetic flux tubes. When the field strength exceeds a critical value, an undulatory instability sets in (Spruit and Ballegooijen, 1982 [48]; Ferriz-Mas and Schüssler, 1993 [49]; Ferriz-Mas and Schüssler, 1995 [50]), which is akin to the Parker instability and favors non-axisymmetric perturbations as sketched in Fig. 8. A downflow of plasma along the field lines within the flux tube leads to an upward buoyancy force acting on the outward displaced parts and a downward force on the troughs, so that the perturbation grows. As a consequence, flux loops form, rise through the convection zone, and finally emerge at the surface to form magnetically bipolar regions and sunspots (Dsilva and Choudhuri, 1993 [51]; Fan et al., 1994 [52]; Caligari et al., 1995 [53]; Fisher et al., 2000 [54]). This is in accordance with the 'rising tree' picture (Zwaan, 1978 [55]) of a partially fragmented magnetic structure, which rises towards the surface and emerges in a dynamically active way. Only later, after the initial stage of flux emergence, the surface fields come progressively under the influence of differential rotation and the convective flow patterns.

The dynamics of such concentrated magnetic structures can be conveniently described using the concept of *isolated magnetic flux tubes*. In ideal MHD we define them as bundles of magnetic field lines (constant magnetic flux), which are separated from the non-magnetic environment by a tangential discontinuity (surface current). As a consequence, the coupling between

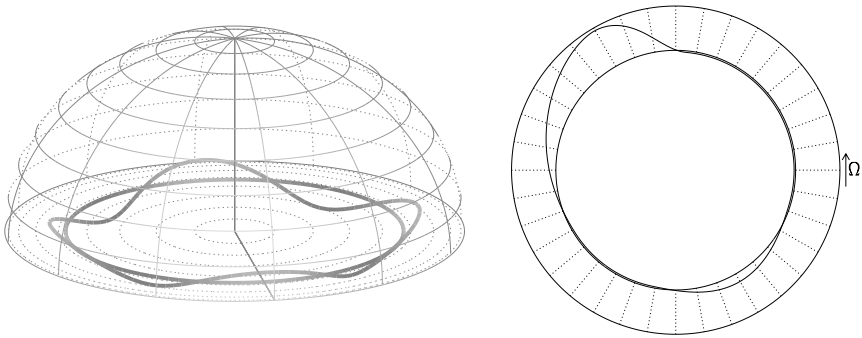


Fig. 8. Left: Undulatory instability of a toroidal flux tube. In the case sketched here, an initially axisymmetric magnetic flux tube (a flux ring) is perturbed by a displacement with azimuthal wavenumber $m = 4$. A downflow of plasmas from the crests into the troughs lets the summits rise while the valleys sink. Right: Polar view of an emerging loop resulting from the nonlinear development of an undulatory instability with azimuthal wave number $m = 2$. Shown is a projection of the tube on the equatorial plane. Note the distinct geometric asymmetry between the two legs of the emerging loop resulting from the Coriolis force.

the tube and its environment becomes purely hydrodynamic and mediated by pressure forces, so that the flux tube can move through a perfectly conducting surrounding plasma, similar to solid body in a fluid.

If the diameter of the flux tube is small compared to all other relevant length scales (scale heights, wavelengths, radius of curvature, etc.) the *thin flux tube approximation* can be employed, a quasi-1D description that greatly simplifies the mathematical treatment (Spruit, 1981 [56]; Ferriz-Mas and Schüssler, 1993 [49]). The forces which are most important for the dynamics of a magnetic flux tube are the *buoyancy force*, the *magnetic curvature force* (if the tube is non-straight), the *Coriolis force* (due to rotation), and the *aerodynamic drag force* (for motion relative to the surrounding plasma). The left panel of Fig. 9 sketches the geometry of these forces for an axisymmetric toroidal flux tube (a flux ring) in a plane parallel to the solar equator. The drag force (not shown) is always anti-parallel to the velocity of the tube relative to the ambient gas, the buoyancy force (anti-)parallel to gravity, while the curvature and rotational (Coriolis and centrifugal) forces are perpendicular to the rotational axis of the Sun.

Rotation has a strong influence on axisymmetric flux tubes rising due to buoyancy (Choudhuri and Gilman, 1987 [57]): the Coriolis force deflects the flux rings to high latitudes unless their initial field strength is larger than about 10^5 G. A simple example may illustrate the effect (right panel of Fig. 9). Assume a flux ring in the equatorial plane, which has been set into an expanding motion with velocity u_r (by the action of a buoyancy force, for example). In a rotating system, this motion causes a Coriolis force which

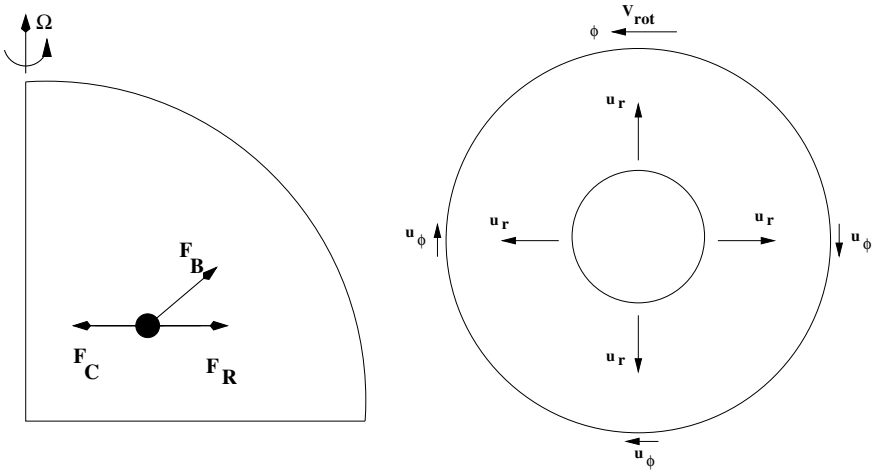


Fig. 9. Left: Forces on an axisymmetric flux tube. The buoyancy force (F_B), rotationally induced force (F_R), and magnetic curvature force (F_C) are shown in a meridional plane. The cross section of the flux tube is indicated by the black disk. While F_R and F_C are perpendicular to the axis of rotation, the buoyancy force is (anti)parallel to the radial direction of gravity. Right: Retarding effect of rotation on an expanding flux ring in the equatorial plane. The rotational counterflow, u_ϕ, in the flux ring driven by the ϕ-component of the Coriolis force leads to an inward directed radial component of the Coriolis force.

drives a flow u_ϕ along the tube and against the direction of rotation, reflecting the conservation of angular momentum. This velocity, in turn, causes an *inward* directed Coriolis force, which acts as a restoring force with respect to the expansion of the flux ring. As a consequence, the radial motion of the flux ring is transformed into an *inertial oscillation* with a period 2Ω, whose amplitude depends on the strength of the initial buoyancy force. If we are outside the equatorial plane, however, the Coriolis force affects only the component of the buoyancy force perpendicular to the axis of rotation (see left panel of Fig. 9); the trajectory of the flux ring then becomes the superposition of an inertial oscillation in the plane perpendicular to the axis of rotation and a rise parallel to the axis. If the Coriolis force dominates (i.e., if the rise time is longer than the period of the inertial oscillation) this leads to the eruption of magnetic flux in high latitudes.

To order of magnitude, the ratio of buoyancy force (F_B) to Coriolis force (F_R) can be estimated as

$$\frac{|\mathbf{F}_B|}{|\mathbf{F}_R|} \simeq \frac{B^2/(8\pi H_p)}{2\rho v_{\rm rise}\Omega} = \left(\frac{B}{B_{\rm eq}}\right)\left(\frac{\rm Ro}{2}\right),$$

where H_p is the pressure scale height, ρ the density, Ω the angular velocity and $B_{\rm eq} = \sqrt{4\pi\rho}\, v_c$ is the equipartition field strength with respect to the convective velocity, v_c. Ro $= v_c/(2H_p\Omega)$ is the *Rossby number*. The rise velocity, $v_{\rm rise}$, of the flux tube has been assumed to be of the order of the Alfvén speed, $v_{\rm A} = B/\sqrt{4\pi\rho}$ (Parker, 1975 [58]). For the lower convection zone of the Sun we have Ro $\simeq 0.2$. Consequently, with $B_{\rm eq} \simeq 10^4$ G, a magnetic field must have a field strength of at least

$$B \;\simeq\; 10\, B_{\rm eq} \;\simeq\; 10^5\,{\rm G}$$

in order to avoid being dominated by the Coriolis force when erupting to form sunspots and active regions. This simple estimate is confirmed by numerical simulations of the rise of magnetic flux tubes (Choudhuri, 1989 [59]). Emerging flux loops are deflected poleward in the same way as rising flux rings unless the initial field strength at the bottom of the convection zone is of the order of 10^5 G. Fields of the same order of magnititude are required to reproduce the observed tilt angles of sunspot groups with respect to the East-West direction (Dsilva and Choudhuri, 1993 [51]; Caligari et al., 1995 [53]) and the asymmetry of their proper motions (Moreno et al., 1994 [60]).

3.3 Global Transport of Magnetic Flux

A large part of the emerged flux is rapidly removed from the solar surface by cancellation with opposite-polarity flux within a timescale of days. The remaining flux becomes redistributed over the solar surface through advection by supergranulation (turbulent diffusion), differential rotation, and meridional flow. The corresponding timescales are in the range of months to years. The evolution of the large-scale patterns of the magnetic flux distribution on these scales (see Fig. 10) can be fairly well reproduced by simulations of passive horizontal flux transport by these flow patterns (Wang and Sheeley, 1994 [61]). In particular, the polarity reversals of the polar caps can be attributed to the dominant transport of opposite-polarity flux from the more poleward following parts of the tilted bipolar regions in the activity belts.

The large-scale evolution of the solar surface flux affects the distribution of regions of open and closed flux in the corona and thus determines the strength and geometry of the interplanetary and heliospheric magnetic field. It is thus possible to connect the amount of open flux from the Sun with the flux emergence in the photosphere (Wang et al., 2000a [62]) and the sunspot number (Solanki et al., 2000 [63]; Solanki et al., 2002 [44]).

3.4 The Solar Cycle and Its Long-Term Modulation

The dominant time scale of variability of solar activity is the 11-year solar cycle, or, if one considers the magnetic polarity reversals, the 22-year magnetic cycle (Hale cycle) of the Sun. The 11-year cycle is most prominently

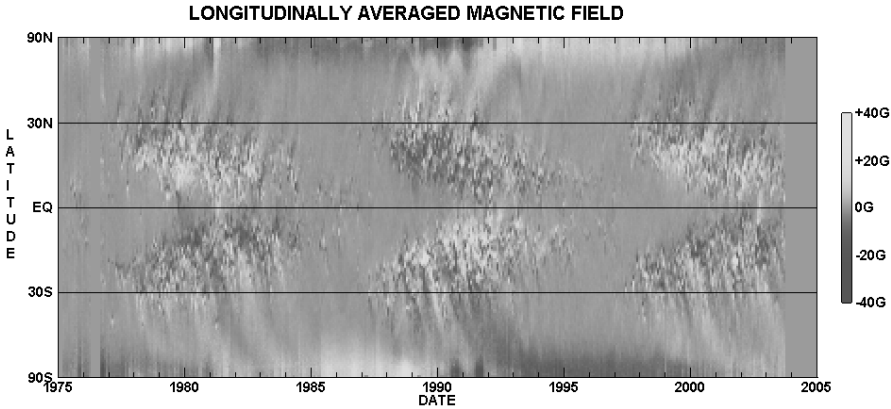


Fig. 10. Time-latitude diagram of the longitudinally averaged magnetic field in the solar photosphere for the last three activity cycles. The emergence of magnetic flux in active regions generates the familiar 'butterfly wings' in lower latitudes. The combined effects of differential rotation, convection, and meridional circulation lead to the magnetic flux transport to high latitudes and thus cause the reversals of the polar magnetic fields in phase with the activity cycle (courtesy D. Hathaway, NASA).

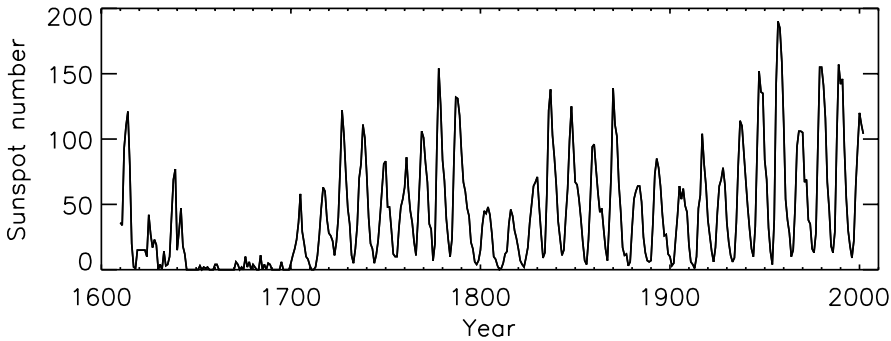


Fig. 11. Relative sunspot number since 1610. Besides the conspicuous 11-year solar cycle, the sunspot record shows century-scale modulation and periods of low activity (*grand minima*). The most prominent of these is the *Maunder minimum* in the 17th century, when only very few sunspots appeared for a period of about 60 years (Eddy, 1976 [68]).

reflected in the record of sunspot numbers starting in the early 17th century (Fig. 11) but many other properties of the Sun also vary with the solar-cycle period. Besides the well-known variations in the magnetic field patterns (e.g., equatorward drift of the sunspot zone, reversal of the polar magnetic fields) and in the frequency of energetic and eruptive events, this includes the total and the spectral radiance (Willson, 1997 [64]; Solanki, 2002 [65]), particularly in the short-wavelength range of the spectrum (UV, EUV, and X-rays), the

shape of the corona, the flux of galactic cosmic rays in the inner heliosphere (Bazilevskaya, 2000 [66]), and also the frequencies of the acoustic eigenmodes of the Sun (Jimènez-Reyes et al., 2001 [67]).

The 11/22-year cycle is modulated by a long-term variability of its amplitude on timescales of decades to centuries. Owing to the unsufficient length of the available data records, it is unclear whether this modulation reflects a superposition of (quasi-)periodic processes or has a chaotic or stochastic character. The strongest perturbation of the sequence of 11-year cycles is the appearance of 'grand minima', during which solar activity is low for an extended (several decades) period of time. The best documented example of a grand minimum is the Maunder minimum between 1640 and 1710, when only very few sunspots where observed (Eddy, 1976 [68]). There are indications that the 11-year cycle nevertheless continued at a low level during the Maunder minimum (Ribes and Nesme-Ribes, 1993 [69]; Beer et al., 1998 [70]; Usoskin et al., 2000 [71]). This was certainly the case for another period of rather low activity, namely the Dalton minimum in the early 19th century with three cycles of low amplitude (see Fig. 11).

On the basis of the varying production rate of the 'cosmogenic' isotopes ^{14}C and ^{10}Be, the long-term modulation of solar activity can be followed further back into the past ((Damon and Sonett, 1991 [72]); Beer, 2000 [73]; Usoskin et al., 2003a [74]). These isotopes are formed by cosmic rays as spallation products in the upper atmosphere of the Earth, from which they are removed by precipitation. The analysis of ^{14}C in tree rings and of the ^{10}Be content in the yearly layers of ice cores drilled in Greenland and Antarctica can be used to determine the variation of the incoming flux of galactic cosmic rays, which itself is anticorrelated with the level of solar activity (Bazilevskaya, 2000 [66]). Consequently, high concentration levels of cosmogenic isotopes indicate low solar activity, and vice versa, so that the sunspot activity during times before the regular telescopic observations can be reconstructed (see Fig. 12).

Magnetic activity of cool stars other than the Sun can be detected indirectly by measuring the associated chromospheric excess emission (Wilson, 1978 [75]). Many stars show cyclic variations similar to the Sun, while others exhibit either irregular variations on a high activity level or have a flat low activity level (Baliunas et al., 1995 [76]). Typically, stars with a cyclic or flat activity level rotate slowly, while those with a high activity level are rapid rotators (Montesinos and Jordan, 1993 [77]). A flat activity level could possibly indicate a grand minimum (Baliunas and Jastrow, 1990 [78]).

4 Origin of the Magnetic Field

The solar activity cycle with its various manifestations and regularities (quasi-periodic variation of the number of sunspots, 22-year magnetic cycle, equatorward drift of the sunspot zone, reversal of the polar magnetic

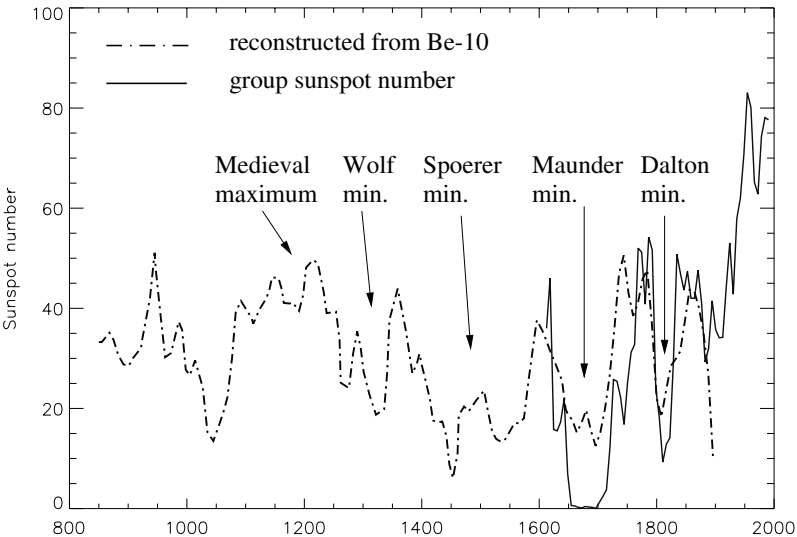


Fig. 12. Directly observed 10-year averaged sunspot numbers (group sunspot number since 1611, solid line) and sunspot numbers reconstructed from the ^{10}Be content in an ice core drilled in Antarctica (dot-dashed line) for the time between AD 850 and 1900 (Usoskin et al., 2000 [71]). Various grand minima and the medieval maximum of solar activity are indicated. The figure clearly shows that the average solar activity since about 1940 is roughly twice as large as the overall average since 850.

fields, etc.) is believed to be the manifestation of a hydromagnetic dynamo mechanism, i.e. the maintainance of a magnetic field against Ohmic decay through the inductive effects of motions in an electrically conducting fluid, operating in the solar convective envelope (Moffatt, 1978 [79]; Parker, 1979 [80]; Ossendrijver, 2003 [81]; Rüdiger and Arlt, 2003 [82]).

4.1 Models of the Solar Dynamo

Three processes are considered to play a major role in the solar dynamo process:

- differential rotation generates toroidal field by shearing poloidal field,
- helical turbulence generates poloidal field from poloidal field, and
- (turbulent) diffusion and meridional circulation distribute, cancel, and transport magnetic flux.

In the simplest case of an axisymmetric (e.g., longitudinally averaged) magnetic field, the poloidal part has field lines in meridional planes (like, e.g., a dipole field) while the field lines of the toroidal field are oriented in the azimuthal (east-west) direction.

The possibility of dynamo action through the combined effects of helical turbulence and differential rotation was first realized by Parker (1955) [83].

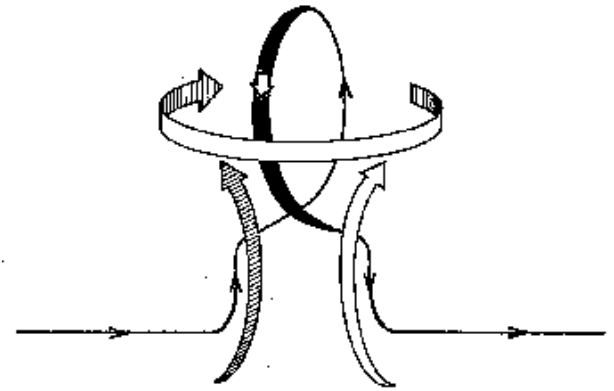


Fig. 13. A rising convective cell expands and rotates through the action of the Coriolis force, twisting a magnetic field line into a loop with components perpendicular to the plane of projection. In this way, a poloidal field component is generated from a toroidal field, and vice versa (after Parker (1979) [85]).

Rising (sinking) convective flows in the stratified, rotating solar convection zone expand (contract) and thus come into helical motion by the action of the Coriolis force. Figure 13 illustrates how such 'cyclonic' motion bends magnetic field lines into twisted loops, which then merge through diffusion and thus form field components perpendicular to the original field. In this way, poloidal field can be generated out of toroidal field, and vice versa.

The effect of small-scale motions on the large-scale magnetic field has been systematically investigated within the framework of mean-field electrodynamics (Krause and Rädler, 1989 [84]). It was formally shown that helical motions drive a mean electric current parallel or antiparallel to the mean magnetic field. This current is represented by a correlation term in the induction equation for the mean magnetic field, called the α-effect. Furthermore, the mean field is subject to enhanced (turbulent) diffusion. The resulting dynamo equations for the mean magnetic field permit periodic wave solutions, which are in accordance with basic phenomena of the solar cycle like the temporally periodic mean field and its migration in heliographic latitude – if the parameters in such models are suitably chosen.

A dynamo process as sketched above is basically 'kinematic': magnetic field lines are passively carried and twisted by the convective flows. As the field strengths grows in the course of the dynamo process the back-reaction of the magnetic field on the generating velocity fields through the Lorenz force becomes important. This limits the field strength that can be reached (at least to order of magnitude) to the equipartition field strength, $B_{\rm eq}$, for which the magnetic energy density equals the kinetic energy density of the generating motions. For the lower half of the solar convection zone we have $B_{\rm eq} \simeq 10^4$ G.

Since, as discussed in Sect. 3.2, the toroidal flux system from which the active regions originate must have a field strength of the oder of 10^5 G, this creates a problem for the conventional theory of turbulent dynamo action. The toroidal field is much stronger than the equipartition value so that convection cannot efficiently twist it to create a poloidal field component via the α-effect. With the convective overshoot region as the seat of this strong toroidal magnetic field, which is wound up by differential rotation in the tachocline (the Ω-effect), three types of dynamo models have been proposed:

- *overshoot layer dynamos* for which both α-effect and differential rotation are located in the overshoot region,
- *interface dynamos* with the Ω-effect working in the overshoot region and the convective α-effect in the turbulent part of the convection zone just above the overshoot layer, and
- *flux transport dynamos* with the α-effect near the surface, originating from the tilt of active regions, and a meridional circulation connecting the separated dynamo processes, with a poleward flow on the surface and an equatorward return flow at the base of the convection zone.

In the case of an overshoot layer dynamo, the regeneration of the poloidal field has to be reconsidered because the strong fields resist the turbulent flow and the kinematic α-effect of convective motions hardly works. Possibilities are magnetic instabilities of magnetic layers (Brandenburg and Schmitt, 1998 [86]; Dikpati and Gilman, 2001b [87]) or helical waves due to the undular instability of magnetic flux tubes, which give rise to a dynamic α-effect (Ferriz-Mas et al., 1994 [88]). The latter instabilities require the field strength to exceed a threshold value. When the field strength falls below the threshold value, dynamo action stops and needs to be 'restarted'. This property may lead to the occurrence of grand minima (see Sect. 4.2).

Interface dynamos Parker (1993) [89] and MacGregor and Charbonneau (1997) [90] employ a spatial separation of the regions where the strong toroidal field is created (in the overshoot layer by differential rotation) and where the weak poloidal field is generated (in the convection zone proper by the conventional α-effect). Both regions are connected by diffusion so that the resulting dynamo wave takes on the character of a surface wave propagating along the interface between the convection zone and the overshoot layer.

In *flux transport dynamos* (Leighton, 1969 [91]; Choudhuri et al., 1995 [92]; Wang and Sheeley, 1991 [93]; Choudhuri et al., 1995 [92]; and Dikpati and Charbonneau, 1999b [94]), the regeneration of the poloidal field from the toroidal field component is assumed to originate with the twist imparted by the Coriolis force on toroidal flux tubes rising through the convective envelope. Thus, the two induction effects, differential rotation and α-effect, are widely separated in space. Dynamo action requires the poloidal field to be transported to the region of strong shear at the bottom of the convection zone. Large-scale meridional circulation in the convection zone, with a poleward surface flow and an equatorward subsurface return flow, is invoked to achieve

this effect. Meridional circulation is observed directly as a poleward flow at the solar surface and helioseismic studies have demonstrated this flow to continue well into the convection zone (Schou and Bogart, 1998 [95]). In flux transport dynamo models, the direction of the meridional (return) flow at the bottom of the convection zone controls the migration of the dynamo wave, which governs the equatorward migration of sunspot belts, while the time period of the dynamo is determined by the flow speed.

4.2 Origin of the Long-Term Modulation

The origin of the long-term modulation of the solar cycle and the 'grand minima' (Sect. 3.4 is hardly understood (Rüdiger, 2000 [96]). In terms of dynamo theory, the modulation of the various dynamo effects has been discussed. Among the mechanism are

- modulation of the differential rotation through the nonlinear back-reaction of the magnetic field,
- stochastic fluctuation of the α-effect,
- variability of the meridional circulation, and
- on-off intermittency due to a threshold for dynamo action.

These effects and their consequences are briefly discussed in the following.

The back-reaction of the large-scale magnetic field on the differential rotation in the tachocline leads to complicated nonlinear behaviour (Weiss and Tobias, 2000 [97]; Moss and Brooke, 2000 [99]; Küker et al., 1999 [100]). Under certain conditions, this includes amplitude modulations and symmetry-breaking bifurcations, so that during periods of weak field (grand minima) magnetic activity dominates in one hemisphere and the symmetry may flip between dipole and quadrupole states (Tobias, 1997b [101]; Beer et al., 1998 [70]).

A different approach to account for the modulation of the solar cycle is the stochastic behaviour of the dynamo excitation owing to, e.g., the finite number of large convective cells (giant cells) in the solar convective envelope (Hoyng, 1987b [102]; Hoyng, 1993 [103]). Mean field dynamo models in which the α-coefficient is a stochastic function of time and latitude show long intervals of low activity reminiscent of grand minima and distinct north-south asymmetries during such periods (Ossendrijver et al., 1996b [104]).

The effect of stochastic variations of the meridional circulation in flux-transport models of the solar cycle is another possible mechanism leading to long-term modulation. Producing extended periods of reduced activity, however, turns out to be rather difficult (Charbonneau and Dikpati, 2000 [105]). On the other hand, the model results suggest that the meridional circulation speed, the primary determinant of the cycle period, acts as a clock, so that the cycle periods rarely depart from their average period for more than a few consecutive periods. The model also exhibits a clear correlation between the toroidal field strength of a given cycle and the strength of the high-latitude

surface magnetic field of the preceding cycle, which is in qualitative agreement with observational inferences (Legrand and Simon, 1981 [106]; Legrand and Simon, 1991 [107]).

The dynamic α-effect due to magnetic buoyancy, one of the candidates for a strong-field dynamo operating in the overshoot layer, only sets in beyond a threshold field strength of several times 10^4 G (Ferriz-Mas et al., 1994 [88]). Therefore, as a starting mechanism, the dynamo requires fluctuating fields, transported by downdrafts from a turbulent convection zone into the overshoot region. On the other hand, such fluctuations, when destructive, can lead to a sequence of low-amplitude cycles or even drive the dynamo subcritical until another, constructive magnetic fluctuation restarts the dynamo again. This leads to on-off intermittent solutions and can be related to the occurrence of grand minima (Schmitt et al., 1996 [98]; Schüssler et al., 1997 [108]). Figure 14 shows a typical results from a numerical simulation of such a dynamo model. Sufficiently strong fluctuations destroy the cyclic behaviour of the overshoot layer dynamo and lead to irregular activity. Such activity is observed in fast-rotating cool stars. On the other hand, stars with low and non-variable magnetic activity may be in a state with only a turbulent convection zone dynamo active.

5 Outlook

Space weather and the multitude of processes by which the Sun affects the Earth and its near space environment are all connected with the solar magnetic field and its variability. Data from space probes like SOHO, ground-based observations, and theoretical studies have shown that the atmosphere of the Sun, which extends from the visible solar limb over the corona deep into the interplanetary space and thus defines the heliosphere, is an extremely complex and dynamical medium comprising a large range of spatial and temporal scales. The solar magnetic field leads to the formation of a rich variety of structures and connects the various layers. Consequently, the solar atmosphere has to be studied and understood as a single coupled system.

The key issue for the future is to understand the physical processes connected with and dominated by the magnetic field. On the one hand, this approach requires measurements within broad ranges of wavelengths (from the extreme ultraviolet to the submillimeter range), of spatial scales (from 100 km, the typical size of photospheric magnetic flux elements, to the solar radius of 700 Mm), and of temporal scales (between dynamical times of seconds up to 11 years, the duration of the activity cycle). On the other hand, such measurements have to be complemented by comparable efforts in MHD theory and simulation.

A better understanding of solar-terrestrial relations requires a 'holistic' view of the Sun and heliosphere as a system connected by the restless magnetic field.

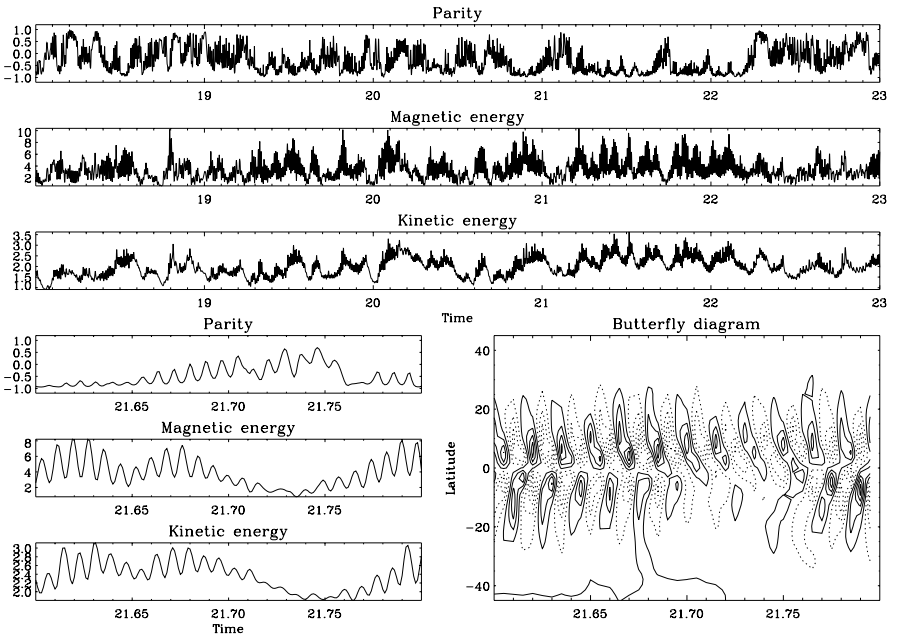


Fig. 14. Time evolution of an on-off-intermittent dynamo with a lower threshold in field strength for dynamo action, a fluctuating source term and, as limiting nonlinearity, a back-reaction of the magnetic field on the differential rotation. The three panels in the upper half give the parity (antisymmetric/dipolar= -1, symmetric/quadrupolar= $+1$) of the magnetic field, the magnetic energy of the toroidal field, and the kinetic energy of the velocity perturbation, respectively (in arbitrary units). In the lower half, the same quantities for a small section of the time series, together with the corresponding butterfly diagram (contour plot of the azimuthal field on a latitude-time plane). The dominant dipolar parity during the oscillatory phases gives way to a mixed parity in 'grand minima'. Typically, a strong north-south asymmetry develops during such phases (Schmitt et al., 1998 [109]).

References

1. M. Stix: *The Sun: an introduction*, 2nd edn (Springer, Berlin Heidelberg New York 2002)
2. C. J. Schrijver and C. Zwaan: *Solar and stellar magnetic activity* (Cambridge University Press 2000)
3. E. R. Priest: *Solar magneto-hydrodynamics*, Geophysics and Astrophysics Monographs (Dordrecht, Reidel 1984)
4. A. N. Cox, W. C. Livingston, and M. S. Matthews (eds.): *Solar interior and atmosphere.* (Tucson, University of Arizona Press 1991)
5. J. T. Schmelz and J. C. Brown (eds.): *The Sun: A Laboratory for Astrophysics* NATO ASIC Proc. 373, (Dordrecht, Kluwer 1992)

6. J. Christensen-Dalsgaard: *Lecture Notes on Stellar Oscillations,* 5th ed. (University of Aarhus, Denmark 2003), `http://astro.phys.au.dk/~jcd/oscilnotes`
7. J. Christensen-Dalsgaard, W. Dappen, S. V. Ajukov et al.: Science, **272**, 1286 (1996)
8. A. G. Kosovichev, J. Schou, P. H. Scherrer et al.: Sol. Phys., **170**, 43 (1997)
9. J. N. Bahcall, S. Basu, and M. H. Pinsonneault: Solar models: structure, neutrinos, and helioseismological properties, In: *The Dynamic Sun* ed. by B. N. Dwivedi (Cambridge University Press 2003), p 8
10. M. Schüssler: Convection. In *The Sun: A Laboratory for Astrophysics*, NATO ASIC Proc. 373 , ed. by J. T. Schmelz and J. C. Brown, (Dordrecht, Kluwer 1992) p 81
11. F. Moreno-Insertis, M. Schüssler, and A. Ferriz-Mas: A&A, **264**, 686 (1992)
12. J. Schou, H. M. Antia, S. Basu et al.: ApJ, **5**05 (1998)
13. S. K. Solanki: Space Sci. Rev., **63**, 1 (1993)
14. U. Narain and P. Ulmschneider: Space Sci. Rev. **54**, 377 (1990)
15. U. Narain and P. Ulmschneider: Space Sci. Rev. **75**, 453 (1996)
16. E. Friis-Christensen, C. Fröhlich, J. Haigh, M. Schüssler, and R. von Steiger (eds.): *Solar Variability and Climate* (Dordrecht, Kluwer 2000)
17. A. Wilson (ed.): *The Solar Cycle and Terrestrial Climate*, ESA SP-463 (Nordwiijk, European Space Agency 2000)
18. A. R. Choudhuri: *The physics of fluids and plasmas : an introduction for astrophysicists* (Cambridge University Press 1998)
19. M. R. E. Proctor and N. O. Weiss: Rep. Prog. Phys. **45**, 1317 (1982)
20. N. E. Hurlburt and J. Toomre: ApJ **327**, 920 (1988)
21. E. N. Parker: ApJ **221**, 368 (1978)
22. H. C. Spruit and E. G. Zweibel: Sol. Phys. **62**, 15 (1979)
23. C. Zwaan, J. J. Brants, and L. E. Cram: Sol. Phys. **95**, 3 (1985)
24. L. R. Bellot Rubio, I. ;. Rodríguez Hidalgo, M. Collados, E. Khomenko, and B. Ruiz Cobo: ApJ **560**, 1010 (2001)
25. P. Venkatakrishnan: Nature **322**, 156 (1986)
26. H. Lin: ApJ **446**, 421 (1995)
27. N. E. Hurlburt, M. R. E. Proctor, N. O. Weiss, and D. P. Brownjohn: J. Fluid Mech. **207**, 587
28. N. O. Weiss, D. P. Brownjohn, P. C. Matthews, and M. R. E. Proctor: Mon. Not. Roy. Astr. Soc. **283**, 1153 (1996)
29. H.C. Spruit, M. Schüssler, and S.K. Solanki: Filigree and flux tube physics. In: *Solar Interior and Atmosphere* ed by A.N. Cox, W.C. Livingston,and M.S. Matthews (Tucson, The University of Arizona Press 1991) p 890
30. K. P. Topka, T. D. Tarbell, and A. M. Title: ApJ **484** (1997)
31. M. Fligge, S. K. Solanki, and Y. C. Unruh: A&A **353**, 380 (2000)

32. Å. Nordlund and R. F. Stein: Solar Magnetoconvection. In *Solar Photosphere: Structure, Convection and Magnetic Fields*, IAU Symp. 138, ed by J. O. Stenflo (Dordrecht, Kluwer Academic Publishers 1990), p 191
33. A. Vögler, S. Shelyag, M. Schüssler, F. Cattaneo, Th. Emonet, and T. Linde: Simulation of solar magneto-convection. In: *Modelling of Stellar Atmospheres*, ed by N. E. Piskunov, W. W. Weiss, and D. F. Gray, San Francisco, Astronomical Society of the Pacific 2003) in press
34. A. Vögler and M. Schüssler: Astron. Nachr./AN **324**, 399 (2003)
35. F. Cattaneo, D. Lenz, and N. Weiss: ApJ **563**, L91 (2001)
36. C. Zwaan and K. L. Harvey: Patterns in the solar magnetic field. In *Solar Magnetic Fields* ed by M. Schüssler and W. Schmidt (Cambridge, Cambridge University Press 1994) p 27
37. V. Gaizauskas, K. L. Harvey, J. W. Harvey, and C. Zwaan: ApJ **265** 1056 (1983)
38. K. L. Harvey and S. F. Martin: Sol. Phys. **32** 389 (1973)
39. K. L. Harvey: The cyclic behavior of solar activity. In *The Solar Cycle*, ASP Conf. Series Vol. 27, ed by K. L. Harvey (San Francisco, Astronomical Society of the Pacific 1992) p 335
40. K. L. Harvey: *Magnetic Bipoles on the Sun*. PhD thesis, University of Utrecht, The Netherlands, 1993.
41. C. J. Schrijver, A. M. Title, A. A. van Ballegooijen, H. J. Hagenaar, and R. A. Shine: ApJ **487**, 424 (1997)
42. H. J. Hagenaar: ApJ **555**, 448 (2001)
43. K. L. Harvey: The solar magnetic cycle. In *Solar Surface Magnetism* ed by R. J. Rutten and C. J. Schrijver, (Dordrecht, Kluwer 1994), p 347
44. S. K. Solanki, M. Schüssler, and M. Fligge: A&A **383**, 706 (2002)
45. M. Schüssler, P. Caligari, A. Ferriz-Mas, and F. Moreno-Insertis: A&A **281**, L69 (1994)
46. F. Cattaneo and D. W. Hughes: J. Fluid Mech. **196**, 323 (1988)
47. Y. Fan: ApJ **546**, 509 (2001)
48. H. C. Spruit and A. A. van Ballegooijen: A&A **106**, 58 (1982)
49. A. Ferriz-Mas and M. Schüssler: Geophys. Astrophys. Fluid Dyn. **72**, 209 (1993)
50. A. Ferriz-Mas and M. Schüssler: Geophys. Astrophys. Fluid Dyn. **81**, 233 (1995)
51. S. D'Silva and A. R. Choudhuri: A&A **272**, 621 (1993)
52. Y. Fan, G. H. Fisher, and A. N. McClymont: ApJ **436**, 907 (1994)
53. P. Caligari, F. Moreno-Insertis, and M. Schüssler: ApJ **441**, 886 (1995)
54. G. H. Fisher, Y. Fan, D. W. Longcope, M. G. Linton, and A. A. Pevtsov: Sol. Phys. **192**, 119 (2000)
55. C. Zwaan: Sol. Phys. **60**, 213 (1978)
56. H. C. Spruit: A&A **102**, 129 (1981)
57. A. R. Choudhuri and P. A. Gilman: ApJ **316**, 788 (1987)
58. E. N. Parker: ApJ **198**, 205 (1975)

59. A. R. Choudhuri: Sol. Phys. **123**, 217 (1989)
60. F. Moreno-Insertis, P. Caligari, and M. Schuessler: Sol. Phys. **153**, 449 (1994)
61. Y. M. Wang and N. R. Sheeley: ApJ **430**, 399 (1994)
62. Y.-M. Wang, J. Lean, and N. R. Sheeley: Geophys. Res. Lett. **27**, 505 (2000)
63. S. K. Solanki, M. Schüssler, and M. Fligge: Nature **408**, 445 (2000)
64. R. C. Willson: Science **277**, 1963 (1997)
65. S. K. Solanki: How does the magnetic cycle change radiance and irradiance of the Sun? In *ESA SP-508: From Solar Min to Max: Half a Solar Cycle with SOHO* (Noordwijk, European Space Agency 2002) p 173
66. G. A. Bazilevskaya: Space Sci. Rev. **94**, 25 (2000)
67. S. J. Jiménez-Reyes, T. Corbard, P. L. Pallé, T. Roca Cortés, and S. Tomczyk: A&A **379**, 622 (2001)
68. J. A. Eddy: Science **192**, 1189 (1976)
69. J. C. Ribes and E. Nesme-Ribes: A&A **276**, 549 (1993)
70. J. Beer, S. Tobias, and N. Weiss: Sol. Phys. **181**, 237 (1998)
71. I. G. Usoskin, K. Mursula, and G. A. Kovaltsov: A&A **354**, L33 (2000)
72. P. E. Damon and C. P. Sonett: Solar and terrestrial components of the atmospheric C-14 variation spectrum. In *The Sun in Time*, p 360 (1991)
73. J. Beer: Space Sci. Rev. **94**, 53 (2000)
74. I. G. Usoskin, S. K. Solanki, M. Schüssler, K. Mursula, and K. Alanko: Phys. Rev. Lett., in press (2003)
75. O. C. Wilson: ApJ **226**, 379 (1978)
76. S. L. Baliunas, R. A. Donahue, W. H. Soon et al.: ApJ **438**, 269 (1995)
77. B. Montesinos and C. Jordan: Mon. Not. Roy. Astr. Soc. **264**, 900 (1993)
78. S. Baliunas and R. Jastrow: Nature **348**, 520 (1990)
79. H. K. Moffatt: *Magnetic field generation in electrically conducting fluids.* (Cambridge, Cambridge University Press 1978)
80. E. N. Parker: *Cosmical magnetic fields: Their origin and their activity.* (Oxford, Clarendon Press; New York, Oxford University Press 1979)
81. M. Ossendrijver: A&AR **11**, 287 (2003)
82. G. Rüdiger and R. Arlt: Physics of the solar cycle. In *Advances in Nonlinear Dynamos* ed by A. Ferriz-Mas and M. Nunez (London, Taylor & Francis 2003) p 147
83. E. N. Parker: ApJ **122**, 293 (1955)
84. F. Krause and K. H. Rädler: *Mean-field magnetohydrodynamics and dynamo theory.* (Oxford, Pergamon Press 1980)
85. E. N. Parker: ApJ **162**, 665 (1970)
86. A. Brandenburg and D. Schmitt: A&A "**338**, L55 (1998)
87. M. Dikpati and P. A. Gilman: ApJ **559**, 428 (2001)
88. A. Ferriz-Mas, D. Schmitt, and M. Schüssler: A&A **289**, 949 (1994)
89. E. N. Parker: ApJ **408**, 707 (1993)
90. K. B. MacGregor and P. Charbonneau: ApJ **486**, 484 (1997)
91. R. B. Leighton: ApJ **156**, 1 (1969)

92. A. R. Choudhuri, M. Schüssler, and M. Dikpati: A&A **303**, L29 (1995)
93. Y.-M. Wang and N. R. Sheeley: ApJ **375**, 761 (1991)
94. M. Dikpati and P. Charbonneau: ApJ **518**, 508 (1999)
95. J. Schou and R. S. Bogart: ApJ **504**, L131 (1998)
96. G. Rüdiger: The dynamo theory for the maunder minimum. In *The Solar Cycle and Terrestrial Climate*, ESA SP-463, ed by A. Wilson (Noordwijk, European Space Agency 2000) p 101
97. N. O. Weiss and S. M. Tobias: Space Sci. Rev. **94**, 99 (2000)
98. D. Schmitt, M. Schüssler, and A. Ferriz-Mas: A&A **311**, L1 (1996)
99. D. Moss and J. Brooke: Mon. Not. Roy. Astr. Soc. **315**, 521 (2000)
100. M. Küker, R. Arlt, and G. Rüdiger: A&A **343**, 977 (1999)
101. S. M. Tobias: A&A **322**, 1007 (1997)
102. P. Hoyng: A&A **171**, 357 (1987)
103. P. Hoyng: A&A **272**, 321 (1993)
104. A. J. H. Ossendrijver, P. Hoyng, and D. Schmitt: A&A **313**, 938 (1996)
105. P. Charbonneau and M. Dikpati: ApJ **543**, 1027 (2000)
106. J. P. Legrand and P. A. Simon: Sol. Phys. **70**, 173 (1981)
107. J. P. Legrand and P. A. Simon: Sol. Phys. **131**, 187 (1991)
108. M. Schüssler, D. Schmitt, and A. Ferriz-Mas: Long-term variation of solar activity by a dynamo based on magnetic flux tubes. In *Advances in Physics of Sunspots*, ASP Conf. Ser. 118 (San Francisco, Astronomical Society of the Pacific 1997) p 39,
109. D. Schmitt, M. Schüssler, and A. Ferriz-Mas. Variability of solar and stellar activity by two interacting hydromagnetic dynamos. In *Cool Stars, Stellar Systems, and the Sun*, ASP Conf. Ser. 154 (San Francisco, Astronomical Society of the Pacific 1998) p 1324

The Application of Radio Diagnostics to the Study of the Solar Drivers of Space Weather

Alexander Warmuth and Gottfried Mann

Astrophysikalisches Institut Potsdam, An der Sternwarte 16, 14482 Potsdam, Germany

Abstract. The application of radio observations to the study of the solar drivers of space weather – flares and CMEs – is reviewed. The different radio emission mechanisms relevant in the solar corona and in interplanetary space are discussed, with an emphasis on plasma emission. The principal types of instrumentation used are presented, as well as some basic techniques which can be applied to extract information from the obtained data. The different kinds of solar radio bursts (intense non-thermal radio emission features observed in dynamic radio spectra) are discussed. They are categorized according to the observing frequency range and with respect to their spectral characteristics. Finally, the different applications of radio data to space weather studies are reviewed: radio observations can be used to study the underlying physics of solar eruptive events, they can be correlated with effects at the Earth in order to forecast space weather hazards, and they allow the study and tracking of potentially geo-effective disturbances (interplanetary coronal mass ejections and shock waves) from the Sun through interplanetary space to the Earth.

1 Introduction

Solar eruptive events – mainly flares and coronal mass ejections (CMEs) – are the main drivers of space weather. A general scheme of how these phenomena can potentially influence the geomagnetic environment of the Earth is shown in Fig. 1 (here we restrict ourselves to particles and bulk mass motions, omitting the various effects due to flare-generated energetic electromagnetic radiation). The geo-effective agents are generated at the Sun, they propagate through the interplanetary (IP) medium, and finally interact with the Earth's magnetosphere.

Shock waves which propagate away from the Sun are launched by flares and/or fast CMEs. A fraction of the shocks (often the ones driven by CMEs) can penetrate into the IP medium (where they are called *IP shocks*) and reach the Earth. At these shocks, particles can be accelerated to high energies, creating *gradual solar energetic particle (SEP) events* which may have severe geo-effective consequences. Alternatively, particles can be directly accelerated in the impulsive phase of a flare, and provided they encounter open magnetic field lines, they can escape and propagate into the IP medium.

A. Warmuth and G. Mann, The Application of Radio Diagnostics to the Study of the Solar Drivers of Space Weather, Lect. Notes Phys. **656**, 51–70 (2005)
http://www.springerlink.com/

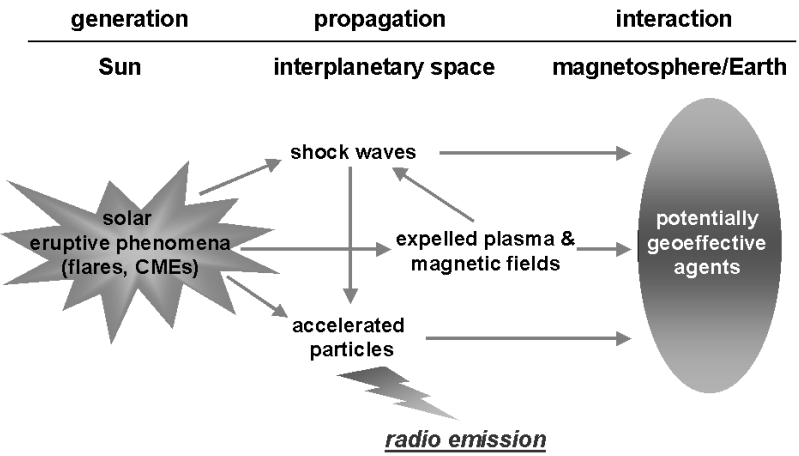


Fig. 1. General scheme of the interactions relevant for space weather: solar eruptive events are produced at the Sun (*left*), the disturbances they create propagate through the interplanetary medium (*middle*) and finally interact with the Earth's magnetosphere (*right*).

These *impulsive SEP events* are more frequent than the gradual ones, but are less important in terms of space weather. For a comprehensive review of both types of SEP events, see Reames (1999) [43].

The matter and magnetic fields that are expelled from the corona by CMEs propagate through the IP medium, where they are called ICMEs or, if measured in-situ, magnetic clouds (Burlaga et al., 1981 [5]). They may interact with the Earth's magnetosphere, potentially producing a geomagnetic storm (which is most likely when the IP magnetic field has a southward component; see e.g. Chen et al. (1997) [8]), the most severe disturbance of the geomagnetic environment (e.g. Gosling, 1993 [19], Webb, 1995 [56]).

As will be seen in Sect. 2, intense solar radio radiation – so-called *solar radio bursts* – can only be generated via a nonthermal mechanism which requires the presence of accelerated electrons. These electrons are either accelerated directly by flares, or by shock waves created by flares or CMEs. Therefore, radio observations provide a particularly valuable tool for studying the solar drivers of space weather and the disturbances they are creating – coronal and IP shocks, energetic particles, and transient disturbances of the IP medium.

It should be noted that space weather is also significantly influenced by a process that is not connected to cataclysmic events at the Sun, but which is instead tied to the large-scale topology of its magnetic field. When fast solar wind streams, which originate from regions where the coronal magnetic field lines are open (coronal holes), overtake the slow solar wind component, co-rotating interaction region (CIRs) are formed. Ahead and behind a CIR, a

pair of shocks is formed in this process, at which particles can be accelerated (see, e.g., Mann et al., 2002 [34] and references therein). This mechanism is responsible for quasi-periodic geomagnetic disturbances, while flares and CMEs cause sporadic – and at times severe – geomagnetic effects. We will concentrate on the latter case.

2 The Physics of Solar Radio Emission

2.1 Emission Mechanisms

The Sun is the brightest radio source in the sky. Three different emission mechanisms contribute to the solar radio flux, and which one is dominating depends on the observing frequency and the solar activity. The first component is *thermal free-free emission*, also known as *thermal bremsstrahlung*. It is generated by free electrons (with a Maxwellian distribution function) which are deflected by the Coulomb electric fields of ions. Thermal emission is dominant in the quiet Sun.

The second important emission mechanism on the Sun is gyro emission, which operates through the spiraling motion of the electrons along magnetic field lines at the electron gyro-frequency,

$$\omega_{\mathrm{ge}} = \frac{eB}{m_{\mathrm{e}}} \,, \tag{1}$$

where e is the elementary charge, m_{e} the electron mass, and B the magnetic field strength in Teslas. Depending on whether the electrons are nonrelativistic, mildly relativistic, or highly relativistic, this mechanism is called *cyclotron*, *gyrosynchrotron*, or *synchrotron emission*, respectively. Of these, only cyclotron and gyrosynchrotron emission are important in the context of solar activity: the former is responsible for the bright coronal emission above sunspots (where B is strong), while the latter dominates the high-frequency bursts associated with solar flares (see Sect. 4.1). Gyrosynchrotron emission is emitted at ω_{ge} and its harmonics, with the contribution of a continuum due to line broadening.

Both thermal bremsstrahlung and gyro emission are *incoherent* processes, which means that the electrons act independently to produce radiation. The observed brightness temperature is then dependent on the kinetic temperature of the particles. However, electrons can also be accelerated in phase, producing photons which are also in phase. This *coherent emission* can have much higher brightness temperatures (up to 10^{15} K). On the Sun, coherent radiation is mainly due to *plasma emission*.

2.2 Plasma Emission

On the Sun, plasma emission is predominantly occurring at frequencies below 1 GHz. The dominance of plasma emission in active events is dramatically

shown by the following comparison: while the thermal emission of the quiet Sun amounts to a flux of 3 sfu (solar flux units, 1 sfu = 10^{-22} W m^{-2} Hz^{-1}) at 40 MHz, a typical radio burst produces a flux of 10^5 sfu at the same frequency.

Let us consider a simple 1D model of an electron embedded in a plasma. In perfect equilibrium, the electric fields of the positive and negative charges will cancel out, but when the electron is displaced from its equilibrium position, it will feel a net electric field E along the x-direction. The force F acting on the electron is given by the equation of motion,

$$F = m_\mathrm{e}\ddot{x} = -eE \,, \tag{2}$$

where m_e is the electron mass and x the displacement from the equilibrium position. The field E is generated by the background distribution of the other charged particles in the plasma, which is represented by the charge density, ϱ. The relation between the two variables is given by Poisson's equation,

$$\mathrm{div}\ \boldsymbol{D} = \varrho \,, \tag{3}$$

where $\boldsymbol{D}$ is the displacement vector. In our model, this simplifies into

$$\frac{dE}{dx} = \frac{eN_e}{\varepsilon_0} \,, \tag{4}$$

where N_e is the total electron number density and ε_0 the permittivity of vacuum. here, the ions are considered to be immobile due to their high mass with respect to the electron mass. Substituting (4) into (2) then gives

$$\ddot{x} + \frac{e^2 N_e}{\varepsilon_0 m_\mathrm{e}} x = 0 \,. \tag{5}$$

Equation (5) describes a classical oscillator. The electron undergoes a harmonic oscillation with the circular frequency

$$\omega_\mathrm{pe} = \sqrt{\frac{e^2 N_e}{\varepsilon_0 m_\mathrm{e}}} \,, \tag{6}$$

which is called the *electron plasma frequency* (often, $f_{pe} = \omega_{pe}/2\pi$ is used instead). This is the natural frequency of electrostatic oscillations of electrons in a plasma.

Plasma emission works in the following way. The plasma is externally disturbed (e.g. by an electron beam or a shock wave) so that the electrons are excited to oscillate. Then, these electrons emit radio waves. In contrast to gyrosynchrotron emission, where the electrons individually generate radio waves, the electrons of the whole plasma *collectively* emit radio waves in the case of plasma emission.

In a plasma, a large variety of wave modes is present. For our purposes, we are especially interested in *Langmuir waves*, which are high-frequency

electron plasma oscillations modified by the thermal motions of the electrons. Langmuir waves have a frequency slightly above f_{pe} and can be excited by suprathermal electrons. Subsequently, Langmuir waves can be converted into escaping electromagnetic waves by scattering at ion density fluctuations and/or by coalescence with low frequency plasma waves (ion acoustic waves). This mechanism generates the fundamental emission near the plasma frequency f_{pe}, while harmonic emission (usually near $f = 2f_{pe}$) arises from the coalescence of two high-frequency electrostatic waves (e.g. Melrose, 1985 [39]). This rather complicated, non-linear process is referred to as plasma emission.

2.3 Derivation of Radio Source Heights and Speeds

The frequency at which plasma emission occurs is directly proportional to the square root of the electron number density, N_e. A convenient form of (6) is

$$f_{\rm pe} = \omega_{\rm pe}/2\pi \approx 9 \times N_e^{1/2} \,, \tag{7}$$

with N_e given in units of m^{-3}. The coronal (and IP) electron density, on the other hand, is a function of height above the solar surface, with N_e monotonically decreasing with height. Each layer in the solar atmosphere therefore corresponds to a characteristic f_{pe}. When we observe a plasma emission feature and apply a suitable coronal electron density model, we can then derive the height of the emission source above the solar surface. For example, frequencies of hundreds of MHz correspond to the low corona, while emission at 20 kHz occurs at one astronomical unit (see Sect. 4.1).

A motion of the source towards, say, greater heights, will then be observed as a drift of the emission towards lower frequencies, since N_e drops with increasing height. The relationship between the drift rate D_f at the frequency f and the radial source velocity v_{source} is given by

$$D_f = \frac{df}{dt} = \frac{f}{2}\frac{1}{N_e}\frac{dN_e}{dt} v_{\rm source} \tag{8}$$

A commonly used coronal density model was derived by Newkirk (1961) [41]:

$$N_e(R) = \alpha\, N_0\, 10^{4.32 R_{\rm S}/R} \,. \tag{9}$$

It gives the electron number density N_e as a function of radial distance R from the solar surface (normalized to the solar radius R_S), with $N_0 = 4.2 \cdot 10^{10}\,\mathrm{m}^{-3}$, and the enhancement factor $\alpha = 1 - 4$ (dependent on whether the burst takes place in the quiet corona or near an active region). The Newkirk model corresponds to a barometric height behavior of the gravitationally stratified corona with a temperature of 1.4 MK (see Koutchmy, 1994 [26]). A heliospheric density model, which is especially useful when dealing with IP phenomena, has been developed by Mann et al. (1999) [33].

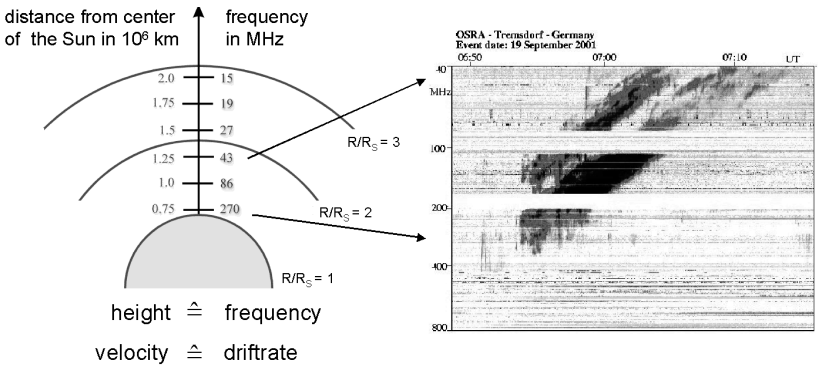


Fig. 2. Schematic of the solar corona (*left*) showing a scale indicating the distance from the center of the Sun (in 10^6 km), and the corresponding plasma frequencies (in MHz) as given by a coronal density model. Using this frequency-height relation, a dynamic radiospectrum (*right*) can be converted to a height-time plot.

Summarizing, each frequency corresponds to a height in the solar atmosphere, and each drift rate to a source velocity. The diagnostic tool which is employed to derive these parameters is the *dynamic radio spectrum*. It represents the intensity of the received radiation in terms of time t and frequency f. In such a spectrum, t runs from left to right, while f decreases from bottom to top, corresponding to increasing height (for a graphic representation, see Fig. 2). Thus, the dynamic radio spectrum represents a path-time diagram of the radio source.

3 Instrumentation and Observational Techniques

The incoming solar radio flux can be characterized by $I = I(f, \theta, \phi, t)$, that is, the intensity I (also known as *flux density*) as a function of frequency f, angular directions θ and ϕ in the plane of the sky, and time t. In addition, the radiation may be polarized, which means that in principle all four Stokes parameters could be defined by f, θ, ϕ, and t. Solar radio telescopes can measure a subset of these parameters of the incoming radiation. For the sake of brevity, let us just consider the total intensity I, bearing in mind that many instruments are capable of measuring additional Stokes parameters (usually V, the degree of circular polarization).

The most basic instrument is the radiometer, which just measures the flux density $I = I(t)$, at a fixed frequency and averaged over the whole Sun. The principal advantages of radiometers are their simplicity and high time cadences.

A radiospectrograph is capable of obtaining dynamic radiospectra $I = I(f, t)$ and could be regarded as a combination of many radiometers at different, closely spaced frequencies. Today, radiospectrographs observing from the

microwave range up to the kilometer regime represent the main asset of solar radio astronomy. A typical example is the Potsdam-Tremsdorf radiospectropolarimeter (Mann et al., 1992 [31]), which measures the solar radio flux in 1024 channels from 800 down to 40 MHz, with a temporal resolution of 0.1 s.

Radioheliographs obtain 2-D radio images of the Sun at one or several fixed frequencies, $I = I(\theta, \phi, t)$, utilizing interferometric techniques. Radioheliograms thus complement the spectral data. Two such instruments which have significantly advanced our knowledge of the radio Sun are the Nançay radioheliograph (Kerdraon and Delouis, 1997 [24]), which observes at five different frequencies in the metric range (164, 237, 327, 411, and 432 MHz), and the Nobeyama radioheliograph (Nakajima et al., 1994 [40]), which operates at two centimeter wavelengths (17 and 34 GHz). Both instruments image the full Sun, the angular resolutions range from 10" at cm wavelengths up to 10' for the metric range.

Finally, it should be noted that ground-based radio observations are only possible above the ionospheric cut-off frequency ($\nu \simeq 7$ MHz), and are already extremely difficult below 20 MHz due to terrestrial interference. Since plasma emission at and below these frequencies originates from IP space, radio instruments have to be space-borne in order to study this regime which is of special importance with respect to space weather. A multitude of spacecraft have been carrying radio receivers, the most important instrument currently being the WAVES radiospectrograph aboard the *Wind* spacecraft (Bougeret et al., 1995 [3]). WAVES covers the dekametric/hectometric/kilometric range from 14 MHz down to 4 kHz.

4 Solar Radio Bursts

We will now discuss the solar radio signatures which are connected with solar eruptive events and which are thus of particular relevance in the context of space weather. Our focus is therefore on the nonthermal solar radio bursts, which are usually generated by plasma emission. Radio bursts are normally classified according to two criteria: the wavelength of observation and the morphological appearance in dynamic radio spectra. These different classification schemes are discussed below, with an emphasis on the underlying physical processes.

4.1 Wavelength Regimes

In order to illustrate the different wavelength ranges, as well as the corresponding source heights, Fig. 3 shows a combined solar radio spectrum from the solar eruptive event of 2 June 2002 (see Classen et al., 2003 [9]).

The three spectra were provided by the Ondřejov radiospectrograph (4 GHz to 800 MHz, see Jiricka et al., 1993 [22]), the Potsdam-Tremsdorf

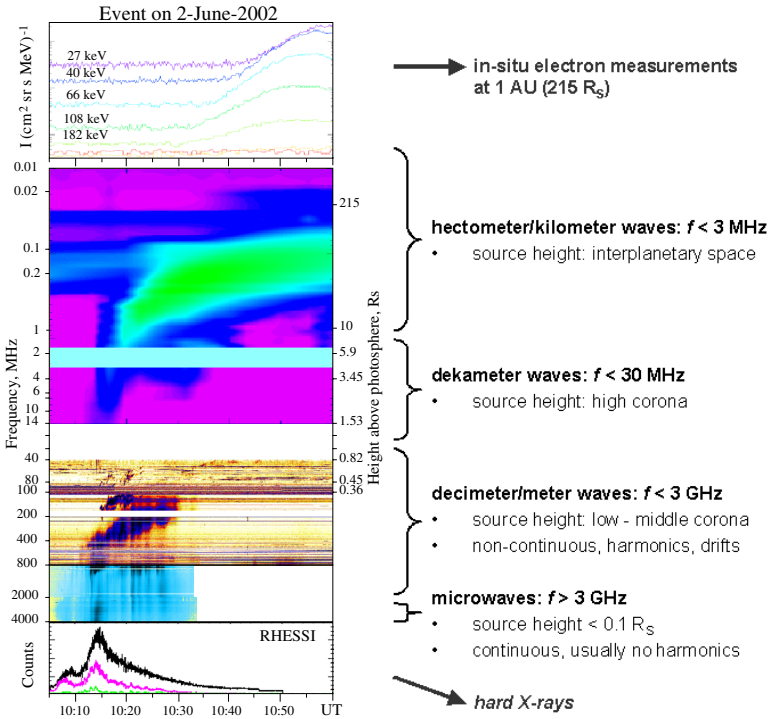


Fig. 3. The solar eruptive event of 2 June 2002 (adapted from Classen et al., 2003 [9]). *Bottom:* hard X-ray count rates (RHESSI). *Middle:* combined dynamic radiospectrum (see main text for details). Frequency decreases from bottom to top (corresponding to increasing height in the solar atmosphere). *Top:* In-situ electron flux measurements (*Wind*/3DP) at different energies. The text to the right of the graph gives some basic information on the radio wavelength ranges.

radiospectropolarimeter (800 MHz to 40 MHz, see Mann et al., 1992 [31]), and by *Wind*/WAVES (14 MHz to 10 kHz, see Bougeret et al., 1995 [3]). In addition, the lowermost and the uppermost parts of Fig. 3 show hard X-ray (HXR) fluxes obtained with the Reuven Ramaty High Energy Solar Spectroscopic Imager (RHESSI, see Lin et al., 2002 [29]) and in-situ measurements of electron fluxes provided by the 3-D Plasma and Energetic Particles Detector (3DP, see Lin et al., 1995 [28]) aboard the *Wind* spacecraft, respectively.

As is the convention for solar radiospectrograms, time runs from left to right and frequency decreases from bottom to top (corresponding to increasing height). In the following, we examine the different spectral regimes, starting at the highest frequencies (i.e. the lowest heights).

Microwaves ($f > 3$ GHz). They originate low in the corona and/or in the chromosphere (at heights of $h < 0.1\ R_S$). Microwave emission is generally broad-band and continuous (i.e. there are no fine structures). Most of the mi-

crowave bursts are due gyrosynchrotron emission by relativistic electrons. Microwave bursts are often closely correlated with the flare's HXR emission (see the HXR lightcurves in Fig. 3), which implies that they are generated by the same energetic electron population that produces the HXR bremsstrahlung – the same particle population that contains the bulk of the energy released in the flare. Therefore, microwave observations of flares provide important information on the primary particle acceleration mechanisms.

Decimeter/Meter Waves ($f < 3$ GHz). This remains the best-studied wavelength regime. Radiation in this range is coming from the low and the middle corona ($h \simeq 1\ R_S$), respectively. In contrast to microwaves, most of the emission is non-continuous, it can be narrow-band, and a multitude of distinct fine structures, harmonics and frequency drifts is observed. The high fluxes and brightness temperatures of the observed bursts require a coherent emission mechanism – namely plasma emission.

Dekameter Waves ($f < 30$ MHz). The dekametric regime is generally similar to the metric, but it originates from the higher corona. Emission mechanisms and morphologies of bursts are also similar. Observations in the dekametric band provide an important link between the comparatively well-known middle corona and the IP space.

Hectometer/Kilometer Waves ($f < 3$ MHz). With respect to emission mechanisms and morphology of bursts, this range is also similar to the meter/dekameter bands. However, as these extremely long wavelengths correspond to the plasma frequency of the IP medium, they are of particular interest in the context of space weather. Observations in this regime allow the tracking of disturbances from the high corona up to the Earth (and even beyond it), which will not be possible with any other kind of observational technique until the Solar Mass Ejection Imager (SMEI; see Radick, 2001 [42]) becomes fully operational.

4.2 Types of Solar Radio Burst

We now proceed to the classification of solar radio bursts according to their appearance in dynamic radiospectra. Note that although this classification was devised for meter-wave bursts, it can be applied to a significantly wider wavelength range (decimetric to kilometric). Figure 4 is an idealized dynamic radio spectrum and shows the basic types of solar radio bursts (as in Fig. 3, frequency decreases from bottom to top). Important characteristics of the bursts are their duration Δt, their bandwidth Δf, and their drift rate, $D_f = df/dt$. In the following, the numerical values given for these parameters will be referring to the metric range. An excellent discussion of the meter-wave bursts can be found in McLean and Labrum (1985) [38].

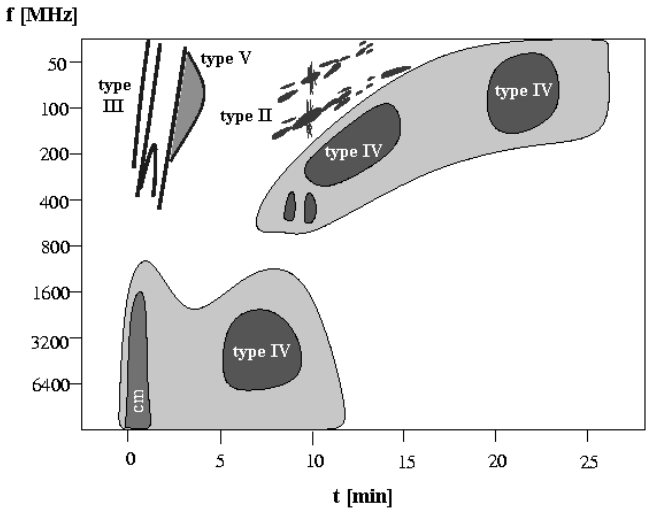


Fig. 4. Dynamic solar radio spectrum showing schematically the basic types of solar radio bursts. Time runs from left to right, frequency decreases from bottom to top (corresponding to increasing height in the solar atmosphere). Time is given in minutes, frequency in MHz.

It is generally accepted that all following bursts are generated by plasma emission (though at least some type IV bursts may be due to gyrosyncrotron emission). The microwave bursts (μ in Fig. 4), which are due to other mechanisms, have already been discussed in Sect. 4.1.

Type I Bursts. Type I bursts (Fig. 5) are characterized by a very short duration (< 1 s), they have bandwidths of a few tens of MHz, and they do not show obvious drifts. Type I bursts are only observed at metric wavelengths and always appear in large numbers, forming irregular structures superposed on a continuous background. These so-called *noise storms* can last for hours to days. Type I emission is therefore not necessarily associated with flares. It is thought to be generated by electrons accelerated to a few thermal energies by an ongoing local energy release in closed coronal structures. Type I bursts are not particularly important for space weather studies, therefore, we will refrain from an in-depth discussion.

Type II Bursts. Type II bursts (Fig. 5) are narrow-band (a few MHz) emission lanes which slowly drift towards lower frequencies ($D_f \simeq 0.1 - 1$ MHz s^{-1}). Both fundamental and harmonic band can be present, and sometimes each band is split into a higher and a lower frequency lane (with a relative separation of $\Delta f/f \simeq 0.1$). For a review, see Mann (1995) [32]. Most bursts are observed in the metric range, but some are also detected in the dekametric

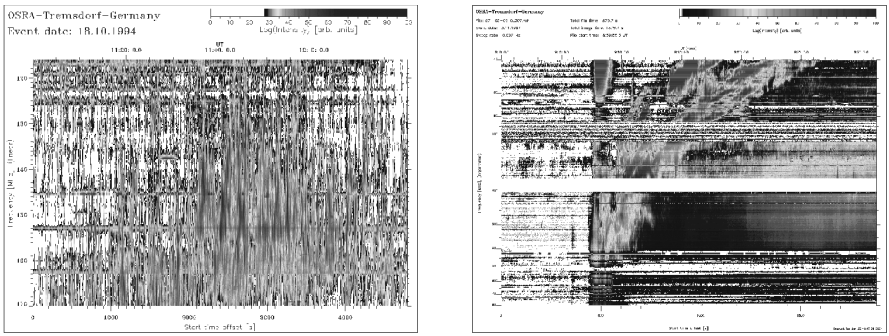


Fig. 5. Dynamic spectra of solar radio bursts (Potsdam-Tremsdorf radiospectropolarimeter). *Left*: A group of type I bursts. Together, they form a noise storm. The individulal bursts are irregularly distributed in frequency and time, no systematic drifts are evident, and the duration of each busrst is very short. *Right*: Type II burst. Note the gradual drift from higher to lower frequencies, the fundamental-harmonic stucture of the two emission lanes, and the band-splitting evident in both of the lanes. The white horizontal bar is a data gap.

to kilometric regimes. These are called IP type II bursts (see, e.g., Cane et al., 1987 [7]).

A type II burst is generated by a magnetohydrodynamic shock wave which propagates outward through the corona. In the corona and in the IP medium, a type II-generating shock is formed when a disturbance exceeds the Alfvén speed

$$v_{\mathrm{A}} = \frac{B}{\sqrt{\mu_0 m_{\mathrm{p}} \mu N}} \,, \tag{10}$$

where μ_0 is the permeability of vacuum, m_p the proton mass, μ the mean molecular weight (0.6 in the corona), and N the total particle number density ($N = 1.92\, N_e$ for $\mu = 0.6$). Velocities of coronal type II sources are of the order of 1000 $\mathrm{km\,s^{-1}}$. At the shock front, electrons are accelerated to suprathermal and/or high energies. They excite Langmuir wave which are then converted into escaping radio waves by the plasma emission process outlined in Sect. 2.2.

Further evidence for electron acceleration is provided by the *herringbone structure* observed in some type II bursts, in which small type III-like bursts (see below) emanate from the "backbone" of the emission lane. These features are interpreted as accelerated electrons which escape from the shock.

Type II bursts are associated both with flares and CMEs, though there is no one-to one correspondence. This has resulted in an extended discussion on the real nature of the shocks which produce the bursts, the candidates being a flare-generated pressure pulse (see e.g. Vršnak and Lulić, 2000a [50] and Vršnak and Lulić, 2000b [51]) or a piston-driven shock created by a CME (Cliver et al., 1999 [11] and references therein). The current view is that both flares *and* CMEs can create shocks (e.g. Classen and Aurass, 2002

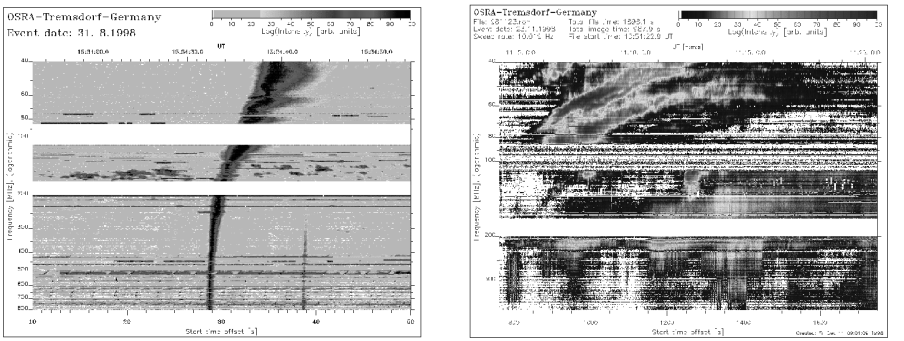


Fig. 6. Dynamic spectra of solar radio bursts (Potsdam-Tremsdorf radiospectropolarimeter). *Left*: Type III burst. It displays a rapid frequency drift and has a short duration. *Right*: Type IV burst (in the lower part of the spectrum). Note the broad-band structure and the slow drift towards lower frequencies. The feature in the upper left of the spectrum is a well-defined type II burst.

[10]; Shanmugaraju et al., 2003 [46]), but it seems that the flare-generated disturbances usually cannot penetrate to IP space, since most of those bursts cease at $\simeq$ 20 MHz (Gopalswamy et al., 1998b [16]). This is probably due to a local maximum of the Alfvén speed in the higher corona (Mann et al., 2003 [35]). Therefore, most hectometric/kilometric type II bursts seem to be generated by CME-driven shocks (Cane et al., 1987 [7]). These bursts are associated with fast CMEs, long-lived energetic solar particle events (Kahler et al., 2003 [23]), and IP shocks, and are therefore particularly relevant for space weather purposes.

Type III Bursts. Type III bursts (Fig. 6) are the most common flare-associated bursts and can occur over a wide frequency range, from $\simeq$ 1 GHz to $\simeq$ 10 kHz, corresponding to a height range extending from the low corona to beyond 1 AU. They are mainly defined by their rapid drift ($D_f \simeq 100\ \mathrm{MHz\,s^{-1}}$) towards lower frequencies, they have a short duration (seconds) and a relatively broad bandwidth ($\Delta f \simeq 100\ \mathrm{MHz\,s^{-1}}$). Many type III bursts display harmonic structure at metric to dekametric wavelengths.

Type III bursts are characteristic of the impulsive phase of solar flares, where they often occur in groups of $\simeq$ 10 bursts, lasting a few minutes. Non-flare associated type IIIs form storm type III bursts, somewhat reminiscent of type I noise storms. The exciting agent of a type III burst is a beam of mildy relativistic electrons ($v \simeq 0.3\ c$) which propagates out of the corona along open magnetic field lines (the beams may also propagate in closed loops, resulting in so-called *inverted-U bursts*). As in the case of type II bursts, the accelerated electrons generate plasma emission. Type III bursts can propagate through IP space up to the Earth, where the radio-generating electrons can be directly observed as impulsive electron events. Type III bursts therefore

give vital clues on the acceleration of electrons in flares, as well as on the propagation of these particles through IP space.

A special class of type III bursts are the so-called *shock-accelerated (SA) type III bursts* (see Cane et al., 1981 [6]; Bougeret et al., 1998 [4]; Klassen et al., 2002 [25]). They start from a type II backbone and are somewhat reminiscent of herringbones, but contrary to them, SA type III bursts extend into the IP medium. Like the herringbones, they are thought to be generated by electron beams which are accelerated at a coronal (or IP) shock.

Type IV Bursts. Type IV bursts are flare-related broad-band continua (Fig. 6). They are divided into two distinct categories: *stationary type IV bursts* show no frequency drift and are characterized as broad-band, long-lasting continuum features which show a wide variety of fine structures – pulsations, zebra patterns and fiber bursts. They follow major flares and may evolve into type I storms. On the other hand, *moving type IV bursts* display a slow drift towards lower frequencies (corresponding to source velocities of up to several 100 $\mathrm{km\,s^{-1}}$), while they are otherwise morphologically similar to stationary type IVs.

Type IV bursts are believed to be either due to plasma emission or due to gyrosynchrotron emission. In any case, the electrons which are responsible for the emission are trapped in a closed magnetic structure. This can be a set of coronal loops (stationary type IV), or a rising structure like an expanding loop or a plasmoid which is ejected during an eruptive event (moving type IV; see, e.g., Stewart, 1985 [49]). The bulk of the electrons remains confined to the magnetic structure due to magnetic mirroring at converging magnetic field lines (i.e., at the feet of coronal loops), therefore, we observe prolonged emission.

Type IV bursts are only seldomly observed in the near-Sun IP medium, but they are nevertheless interesting for solar-terrestrial studies since they can provide valuable information on the energy release mechanism of solar eruptive events. Several flare models require the formation and ejection of plasmoids, and CME cores might actually be sources of type IV bursts.

Type V Bursts. Type V bursts are continuum bursts which start during or immediately after a group of type III bursts. They are possibly created by electrons which have been removed from the type III-generating beam by pitch angle scattering. For the purposes of solar-terrestrial studies, type V bursts are not important.

4.3 An Example – The Solar Eruptive Event of 2 June 2002

We return to Fig. 3 in order to show how observations in different wavelength ranges can be combined to track different radio sources from the lower corona to 1 AU. The event of 2 June 2002 was associated with a moderate C8/SF

flare at the solar coordinates S20W61. This location near the west limb means that the event was magnetically well connected to the Earth.

The lowermost panel in Fig. 3 shows the count rate of HXR photons as measured by the RHESSI instrument (Lin et al., 2002 [29]) in the energy bands of 6-12, 12-25, and 25-50 keV (the higher the energy, the lower is the count rate). Radiation at these energies is primarily due to bremsstrahlung emitted by electrons which precipitate onto the chromosphere.

In the dynamic radio spectrum from 4 GHz to 800 MHz, several fast-drifting microwave bursts are visible. A closer examination shows that the bursts are temporally associated with the peaks of the HXR flux, which means that the radio-generating electrons are accelerated at the same site and by the same mechanism as the bremsstrahlung-emitting electron population is (see Classen et al., 2003 [9]).

In the range from 800 to 40 MHz, several radio bursts can be discerned: the big dark feature drifting from high to low frequencies is a moving type IV burst, while the smaller drifting feature at ~ 150 MHz is the harmonic band of a type II burst. Just above it, the fundamental band is faintly visible (at ~ 80 MHz). At $\sim$10:14 UT, a series of weak type III bursts is observed (beginning at ~ 50 MHz). At lower frequencies (below 14 MHz), the type III bursts, having merged into a single burst, have become the dominating feature in the radiospectrum. From these observations, combined with Nançay radioheliograms, it was concluded that the type II-generating shock moved ahead of the type II source (presumably a plasmoid) at roughly the same velocity ($\sim 500\ \mathrm{km\,s^{-1}}$).

From the in-situ electron flux measurements (top panel in Fig. 3), it was found that the particles were injected at the Sun after both the primary energy release and the the shock wave development. However, there is a good temporal association with minor HXR peaks. It is concluded in Classen et al. (2003) [9] that the high-energy electrons measured at the Earth are accelerated by an ongoing reconnection process in the low corona.

5 Applications of Radio Observations

After having discussed how radio emission is generated and which different kinds of signatures the active Sun creates, we will now show how this knowledge can be employed for the benefit of space weather studies. Basically, three different approaches are possible: radio observations can be used to study the underlying physics of solar eruptive events, they can be correlated with effects at the Earth in order to forecast space weather hazards, and they allow the study and tracking of potentially geo-effective disturbances from the Sun through the IP medium to the Earth.

5.1 Studying the Nature of Flares and CMEs

Radio observations can be used to study the basic nature of flares and CMEs, which includes the primary energy release and particle acceleration. Though this is not directly connected with space weather, it is nevertheless crucial, for without an understanding of the basic drivers of space weather, the reliability of any forecasts remains doubtful.

Since this approach is mainly concerned with the origin and initial development of flares and CMEs, its objects of study are close to the Sun, and consequently observations at the shorter wavelengths are required. Microwave observations can provide many important inferences on energy release and particle acceleration in flares, especially when combined with HXR data). For example, gyrosynchrotron emission from flares can be used to derive coronal magnetic field strengths (see, e.g., Gary and Hurford, 1994 [13]).

With regard to the metric regime, several possibilities of analyzing radio bursts have already been mentioned in Sect. 4.2. Type III bursts offer us very sensitive diagnostics of coronal energy release and particle acceleration. Type II and type IV bursts can be used to study the propagation of coronal shocks (e.g. Vršnak et al., 2002a [53]) and closed magnetic structures (i.e. CME cores or plasmoids, e.g. Gopalswamy et al., 1990 [14]), respectively, and can be used to derive important physical parameters of the corona. For example, if we accept that the band split observed in some type II bursts is due to emission from ahead of and behind the shock front (e.g. Smerd et al., 1975 [47]), then the amount of band split yields the compression factor $X = \varrho_d/\varrho_u$ (where ϱ_u is the upstream and ϱ_d the downstream density) at the shock. Using a simple model (Vršnak et al., 2002b [54]), we can then deduce the Alfvénic Mach number M_A of the shock and, taking the type II speed derived with a coronal density model, we get the Alfvén speed v_A and the magnetic field strength B of the ambient medium. The knowledge of these coronal parameters is vital for the understanding of how a CME develops, and how a coronal shock may penetrate into the IP medium.

Complementary to these spectral techniques, radioheliographic observations offer important inferences on the spatial and kinematic evolution of space weather agents. CMEs can be detected due to the non-thermal particles accelerated during the magnetic energy release. This is possible out to a few solar radii. In radioheliograms, CMEs can be observed also in front of the solar disk, which is not possible with coronagraphs. In addition, a much higher time cadence is possible with radioheliographs. Therefore, a more complete picture of the geometry of the CME and its trajectory is obtained (Kundu et al., 1989 [27]; Maia et al., 2000 [30]; Bastian et al., 2001 [2]; Manoharan et al., 2001 [37]), which is in turn vital for predicting the impact of the CME on the IP medium and, ultimately, on the Earth. The same technique can be applied to study the trajectories of type II, type III and type IV burst sources. One major result of these studies was that most type II bursts have clearly

non-radial (with respect to the normal on the solar surface) trajectories (e.g. Stewart, 1984 [48]; Aurass et al., 1998 [1]).

On the other hand, radioheliographs operating in the microwave range can observe CME-associated prominence eruptions and arcade formation (see Hanaoka et al., 1994 [20]; Gopalswamy et al., 1998a [15]). All these phenomena are important to understand the mechanisms which lead to the launch of a CME. Note also that there are several additional phenomena that are associated with flares and/or CMEs, such as smaller-scale ejecta (e.g. flare sprays, see Vršnak, 2001 [52]) or Moreton waves (e.g. Warmuth et al., 2001 [55]). On the one hand, such phenomena can be sources of shocks and may accelerate particles, while on the other hand they can offer diagnostic tools for studying the underlying physics of solar eruptions.

5.2 Using Radio Events as Predictors of Space Weather Hazards

Apart from being used to study the basic physics of solar eruptions, radio observations can directly be employed as a tool of space weather prediction. The principle here is to look for correlations between radio phenomena and effects at the Earth. The fast CMEs that drive IP shocks and produce SEP events are generally associated with metric/dekametric type II and type IV bursts, and we can use this correlation to predict the arrival of such geo-effective disturbances. While the correlation is quite coarse, which certainly does not allow straightforward prediction of potential geo-effectiveness, it is nevertheless an additional channel of information for space weather forecasting, and combined with other data sets (particularly optical, extreme ultraviolet and soft X-ray imaging) it has proven its usefulness.

As a recent result of such studies, it has been found that CMEs associated with dekametric/hectometric type II bursts are faster and also wider than the average (Gopalswamy et al., 2000 [17]). This is a particularly good indicator of geo-effectiveness.

5.3 Studying and Tracking Interplanetary Disturbances

The final approach of using radio data for space weather purposes is more direct: we can directly observe the disturbances which may become geo-effective as they propagate away from the Sun, through the IP medium, towards the Earth. Here, low-frequency observations that require radio receivers aboard spacecraft are employed. It should be noted that this is presently the only way to track such disturbances between the high corona, where they may still be observable with a coronagraph, and 1 AU. IP disturbances may be measured in-situ, but that is usually possible only near the Earth, and even then only at a few points at best.

IP type III bursts are the most common feature in the long-wavelength regime. While the electrons producing them can reach the Earth, they are not energetic enough to cause significant space weather effects. However, they

offer very important diagnostics of the propagation of particles through the IP medium. As an example, Reiner et al. (1998a) [44] used the radiospectrographs aboard the widely separated Ulysses and Wind spacecraft to reconstruct the 3-D trajectory of an IP type III burst. With these observations, it was also possible to measure the IP plasma density along the path of the radio source.

In terms of relevance to space weather, the study of type II bursts offers us the most interesting and extensive information. Fast CMEs (with velocities of up to 2000 $\mathrm{km\,s^{-1}}$) are highly associated with both SEP events and IP shocks (see Sect. 1), and consequently, with IP type II bursts (e.g. Gopalswamy et al., 2000 [17]). Dekametric/hectometric/kilometric observations with space-based radiospectrographs allow the identification of shock-driving CMEs, and the tracking of the IP type II emission lanes from the high corona up to 1 AU. Combined with a heliospheric density model (e.g. Mann et al., 1999 [33]), this yields the distance between shock and Sun and the shock's velocity (Reiner et al., 1998b [45]). These parameters are a vital input to the models which are used to predict the arrival time of IP shocks and their associated CMEs (see, e.g., Fry et al., 2003 [12]).

Another interesting discovery was that when a fast CME overtakes a slower one, an interaction takes place which shows up as an intense radio continuum superposed on the type II lane (Gopalswamy et al., 2001 [18]). This enhancement is interpreted as a consequence of shock strengthening. The CME interaction has important implications for space weather: since the trajectories of the interacting CMEs can change significantly, this process might be responsible for many false alarms when predicting CME or shock arrivals at the magnetosphere.

A totally different method of tracking ICMEs is the interplanetary scintillation (IPS) technique (Hewish et al., 1964 [21]). It exploits the scattering of radiation from pointlike radio sources (extragalactic objects, i.e. quasars) by small-scale density inhomogeneities in the IP medium. Thus, it is possible to track the compression region (which is associated with enhanced turbulence) between the IP shock which propagates in front of the ICME and the ICME body. The IPS technique provides the speed and the level of density fluctuations of the disturbance, while the observation of a large number of radio sources provides a good spatial resolution. For a recent application of the technique, see Manoharan et al. (2000) [36].

6 Conclusion

Solar non-thermal radio radiation is a valuable tool for studying solar activity, in particular with respect to the solar eruptive phenomena (flares and CMEs) that are the main drivers of space weather. Observations at shorter wavelengths (microwaves to meter waves) provide information on the primary energy release and particle acceleration in flares, as well as on the initial

development of CMEs and smaller-scale ejecta (i.e. plasmoids) and associated phenomena, such as shock waves. Type III bursts (electron beams) are sensitive signatures of energy release and particle acceleration in the corona. Type II bursts track shock waves that are generated by flares (pressure pulse) and/or CMEs (piston mechanism). In summary, radio obseravtions at shorter wavelengths are employed to study the basic physics of the main solar driver of space weather, as well as the phenomena associated with these eruptive events (i.e. coronal waves and shocks, ejecta, particle beams)

At longer wavelengths (dekameter to kilometer waves), the whole height range from the upper corona, through the IP medium, up to and beyond the Earth can be observed. It is in this range where the main causes of geomagnetic disturbances develop and propgagate towards the Earth: fast CMEs drive IP shock waves, which are believed to be primarily responsible for the SEP events that can lead to severe geo-effective consequences. On the other hand, the CMEs themselves, upon reaching the Earth, can interact with the terrestrial magnetosphere, leading to a geomagnetic storm. Using space-borne radio receivers operating in the dekameter to kilometer wavebands, these disturbances can be tracked from the high corona up to the Earth.

As our technical civilization becomes increasingly vulnerable to space weather hazards, the reliability of forecasts has to be increased significantly. In this process, radio data will be a major contribution. A new generation of solar radio instruments is currently under study and/or development. Two projects for ground-based observatories – the Frequency-Agile Solar Radiotelescope(FASR) and the Low Frequency Array (LOFAR) – combine spectralresolution with imaging capability and high time cadence. With these assets, substantial advances in our understanding of solar eruptive events and their consequences for the Earth are to be expected.

Acknowledgements

The work of A. Warmuth was supported by DLR under grant No. 50 QL 0001. We thank H.-T. Classen for the preparation of figures. The data and software for the hard X-ray and particle observations were provided by the RHESSI and Wind/3DP Investigation teams (R.P. Lin, PI) at the University of California, Berkeley. The radio spectrum from the Ondřejov Obseravtory is by courtesy of M. Karlický.

References

1. Aurass, H., Hofmann, A., Urbarz, H.-W.: Astron. Astrophys. **334**, 289 (1998)
2. Bastian, T. S., Pick, M., Kerdraon, A., Maia, D., Vourlidas, A.: Astrophys. J. **558**, L65 (2001)

3. Bougeret, J.-L., Kaiser, M. L., Kellogg, P. J., et al.: Space Sci. Rev. **71**, 231 (1995)
4. Bougeret, J.-L., Zarka, P., Caroubalos, C., et al.: Geophys. Res. Lett. **25**, 2513 (1998)
5. Burlaga, L. F., Sittler, E., Mariani, F., Schwenn, R.: J. Geophys. Res. **86**, 6673 (1981)
6. Cane, H. V., Stone, R. G., Fainberg, J., et al.: Geophys. Res. Lett. **8**, 1285 (1981)
7. Cane, H. V., Sheeley, N. R., Jr., Howard, R. A.: J. Geophys. Res. **92**, 9869 (1987)
8. Chen, J., Cargill, P. J., Palmadesso, P. J.: J. Geophys. Res. **102**, A7-14701 (1997)
9. Classen, H. T., Mann, G., Klassen, A., Aurass, H.: Astron. Astrophys. **409**, 309 (2003)
10. Classen, H. T., Aurass, H.: Astron. Astrophys.**384**, 1098 (2002)
11. Cliver, E. W., Webb, D. F., Howard, R. A.: Solar Phys. **187**, 89 (1999)
12. Fry, C. D., Dryer, M., Smith, Z., Sun, W., Deehr, C. S., Akasofu, S.-I.: J. Geophys. Res. **108**, SSH 5-1/2002JA009474 (2003)
13. Gary, D. E., Hurford, G. J.: Astrophys. J. **420**, 903 (1994)
14. Gopalswamy, N., Kundu, M. R.: Astrophys. J. **365**, L31 (1990)
15. Gopalswamy, N., Hanaoka, Y.: Astrophys. J. **498**, L179 (1998)
16. Gopalswamy, N., Kaiser, M. L., Lepping, R. P., et al.: J. Geophys. Res. **10**, 307 (1998)
17. Gopalswamy, N., Kaiser, M. L., Thompson, B. J., et al.: Geophys. Res. Lett. **27**, 1427 (2000)
18. Gopalswamy, N., Yashiro, S., Kaiser, M. L,; Howard, R. A., Bougeret, J.-L.: Astrophys. J. **548**, L91 (2001)
19. Gosling, J. T.: J. Geophys. Res. **98**, 18937 (1993)
20. Hanaoka, Y., Kurokawa, H., Enome, S., et al.: PASJ **46**, 205 (1994)
21. Hewish, A., Scott, P. F., Wills, D.: Nature **203**, 1214 (1964)
22. Jiricka, K., Karlicky, M., Kepka, O., Tlamicha, A.: Solar Phys. **147**, 203 (1993)
23. Kahler, S. W., Reames, D. V.: Astrophys. J. **584**, 1063 (2003)
24. Kerdraon, A., Delouis, J.-M.: The Nançay Radioheliograph. In: *Coronal Physics from Radio and Space Observations*, edited by Trottet, G., Lect. Notes Phys. **483**, 192 (Springer, Berlin Heidelberg New York 1997)
25. Klassen, A., Bothmer, V., Mann, G., Reiner, M. J., Krucker, S., Vourlidas, A., Kunow, H.: Astron. Astrophys. **385**, 1078 (2002)
26. Koutchmy, S.: Adv. Space Res. **14 (4)**, 29 (1994)
27. Kundu, M. R., Schmahl, E. J., Gopalswamy, N., White, S. M.: Adv. Space Res. **9 (4)**, 41 (1989)
28. Lin, R. P., Anderson, K. A., Ashford, S., et al.: Space Sci. Rev. **71**, 125 (1995)
29. Lin, R. P., Dennis, B. R., Hurford, G. J., et al.: Solar Phys. **210**, 3 (2002)
30. Maia, D., Pick, M., Vourlidas, A., Howard, R.: Astrophys. J. **528**, L49 (2000)

31. Mann, G., Aurass, H., Voigt, W., Paschke, J.: in Proc. 1st SOHO Workshop, ESA SP-348, 129 (1992)
32. Mann, G.: Theory and Observations of Coronal Shock Waves. In: *Coronal Magnetic Energy Releases*, edited by Benz, A., Krüger, A., Lect. Notes Phys. **444**, 183 (Springer, Berlin Heidelberg New York 1995)
33. Mann, G., Jansen, F., MacDowall, R. J., Kaiser, M. L. Stone, R. G.: Astron. Astrophys. **348**, 614 (1999)
34. Mann, G., Classen, H. T., Keppler, E., Roelof, E. C.: Astron. Astrophys. **391**, 749 (2002)
35. Mann, G., Klassen, A., Aurass, H., Classen, H. T.: Astron. Astrophys. **400**, 329 (2003)
36. Manoharan, P. K., Kojima, M., Gopalswamy, N., Kondo, T., Smith, Z.: Astrophys. J. **530**, 1061 (2000)
37. Manoharan, P. K., Tokumaru, M., Pick, M., et al.: Astrophys. J. **559**, 1180 (2001)
38. McLean, D. J., Labrum, N. R.: *Solar Radiophysics*, (Cambridge Univ. Press, Cambridge 1985)
39. Melrose, D. B.: Plasma emission mechanisms. In: *Solar Radiophysics*, edited by McLean, D. J., Labrum, N. R. (Cambridge Univ. Press, Cambridge 1985), pp 177–210
40. Nakajima, H., Nishio, M., Enome, S., et al.: Proc. IEEE **82**, 705 (1994)
41. Newkirk, G. A.: Astrophys. J. **133**, 983 (1961)
42. Radick, R. R.: Proc. SPIE **4498**, 84 (2001)
43. Reames, D. V.: Space Sci. Rev. **90**, 413 (1999)
44. Reiner, M. J., Fainberg, J., Kaiser, M. L., Stone, R. G.: J. Geophys. Res. **103**, 1923 (1998)
45. Reiner, M. J., Kaiser, M. L., Fainberg, G. J., Stone, R. G.: J. Geophys. Res. **103**, 29651 (1998)
46. Shanmugaraju, A., Moon, Y.-J., Dryer, M., Umapathy, S.: Solar Phys. **215**, 161 (2003)
47. Smerd, S. F., Sheridan, K. V., Stewart, R. T.: Astrophys. J. **16**, 23L (1975)
48. Stewart, R.T.: Solar Phys. **94**, 379 (1984)
49. Stewart, R.T.: Moving Type IV Bursts. In: *Solar Radiophysics*, edited by McLean, D. J., Labrum, N. R. (Cambridge Univ. Press, Cambridge 1985), pp 361–383
50. Vršnak, B., Lulić., S.: Solar Phys. **196**, 157 (2000)
51. Vršnak, B., Lulić., S.: Solar Phys. **196**, 181 (2000)
52. Vršnak, B.: J. Geophys. Res. **106**, 25249 (2001)
53. Vršnak, B., Aurass, H., Magdalenić, J., Goplswamy, N.: Astron. Astrophys. **377**, 321 (2002)
54. Vršnak, B., Magdalenić, J., Aurass, H., Mann, G.: Astron. Astrophys. **396**, 673 (2002)
55. Warmuth, A., Vršnak, B., Aurass, H., Hanslmeier, A.: Ap. J. **560**, L105 (2001)
56. Webb, D. F.: Rev. Geophys. Suppl. **33**, 577 (1995)

Interplanetary Disturbances

Robert F. Wimmer-Schweingruber

Institut für experimentelle und angewandte Physik, Extraterrestrische Physik, Universität Kiel, Leibnizstr. 11, 24118 Kiel, Germany

Abstract. The Sun emits the variable solar wind which interacts with the very local interstellar medium to form the heliosphere. Hence variations in solar activity strongly influence interplanetary space, from the Sun's surface out to the edge of the heliosphere. Superimposed on the solar wind are mass ejections from the Sun and/or its corona which disturb the interplanetary medium – hence the name "interplanetary disturbances".

Interplanetary disturbances are the sources of large-scale particle acceleration, of disturbances in the Earth's magnetosphere, of modulations of galactic cosmic rays, in short, they are the prime focus for space weather studies.

This lecture will give an overview of the relevant physical background including magnetic reconnection, particle acceleration, cosmic-ray modulation, as well as an overview of the properties of the solar wind and of interplanetary manifestations of coronal mass ejections – the "interplanetary disturbances".

1 Introduction

Space weather is largely due to the interaction of the changing solar conditions, their interplanetary manifestations, the heliosphere, and the surrounding interstellar and galactic environment. Space-weather studies are environmental studies, albeit not in quite as provincial a fashion as this term is normally used. This section will try to link solar phenomena to their interplanetary consequences and their influence on the interaction of the heliosphere with the galactic environment.

The Sun exhibits a tremendous variability when we consider other properties than simply the "solar constant". Its magnetic structure changes in a cyclic manner, due to the inherently chaotic nature of the dynamo generating its magnetic field. This manifests itself in the overall magnetic structure of the solar atmosphere, the corona, and is convected out into the heliosphere by the solar wind. Intermittent disturbances, which may be important for the removal of magnetic flux of the old polarity over the magnetic activity cycle, are important agents for space weather as it influences mankind.

So-called coronal mass ejections are the most spectacular manifestations of the solar activity cycle. Their onset is discussed in some detail in Sect. 2, where we consider magnetic reconnection, the ultimate releasing agent for these giant solar eruptions. Remote observations of these coronal mass ejections (CMEs) are discussed in Sect. 3. As these ejections move into interplanetary space, they interact with it in various ways, the most important here is

R.F. Wimmer-Schweingruber, Interplanetary Disturbances, Lect. Notes Phys. **656**, 71–129 (2005)
http://www.springerlink.com/

the driving of shocks through the heliosphere and their ultimate fate as global merged interaction regions in the outer heliosphere (Sect. 4). The CME-driven shocks are giant particle accelerators which can accelerate heliospheric particles up to high energies (hundreds of MeV/amu), particle acceleration and other transport processes are considered in Sect. 5. Space weather is also related to space climate. Long-term variations of solar activity, but also in the galactic environment can influence the very local space weather in the Earth's environment (Sect. 6).

2 Magnetic Reconnection and the Ejection of Coronal Mass

2.1 Preliminary Considerations

In order to understand the properties of the plasma and magnetic field in interplanetary space one often uses the so-called magneto-hydro-dynamic (MHD) approximation. It approximates a plasma as a globally electrically neutral but magnetized fluid in which temporal and spatial disturbances are slow or large-scaled, respectively, when compared to the characteristic properties of the plasma such as plasma and cyclotron frequencies and Debye radius. The equations of MHD consist of an equation of continuity for both the mass density and the current density, an equation of motion, Ohm's law, as well as Maxwell's equations. Often the displacement current $\dot{\boldsymbol{E}}/c^2$ can be neglected.

We combine Ampéres and Ohms laws to obtain

$$\boldsymbol{J} = \frac{1}{\mu_0}\boldsymbol{\nabla} \times \boldsymbol{B} = \sigma\left(\boldsymbol{E} + \boldsymbol{u} \times \boldsymbol{B}\right), \tag{1}$$

where σ is the electrical conductivity,

$$\sigma = \frac{ne^2}{m_e}\tau_c, \tag{2}$$

where τ_c is the average collision time in the plasma. We take the curl of (1) and use Faraday's law

$$\dot{\boldsymbol{B}} = -\boldsymbol{\nabla} \times \boldsymbol{E} \tag{3}$$

to obtain,

$$\dot{\boldsymbol{B}} = \boldsymbol{\nabla} \times \boldsymbol{u} \times \boldsymbol{B} + \frac{1}{\mu_0 \sigma}\Delta\boldsymbol{B}, \tag{4}$$

where we have used $\nabla \times \nabla \times \boldsymbol{B} = \nabla(\nabla \cdot \boldsymbol{B}) - \Delta\boldsymbol{B} = -\Delta\boldsymbol{B}$. Equation (4) is the induction equation of MHD. Understanding its structure helps us understand the properties of the interplanetary medium. Depending on the electric conductivity of the plasma, either one of the two terms on the right-hand side

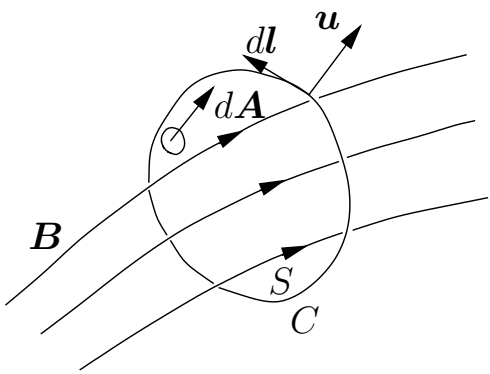


Fig. 1. Magnetic flux through the surface S spanned by the closed curve C.

will dominate. For small conductivity, or for slow movements in the plasma, the induction equation turns into a diffusion equation,

$$\dot{\boldsymbol{B}} = \frac{1}{\mu_0 \sigma} \Delta \boldsymbol{B}, \tag{5}$$

because this term dominates the other term in this case. The quantity $1/\mu_0\sigma$ is called magnetic diffusivity. We can use (5) to derive a diffusion time,

$$\left[\frac{\partial \boldsymbol{B}}{\partial t}\right] = \left[\frac{1}{\mu_0 \sigma} \frac{\partial^2 \boldsymbol{B}}{\partial r^2}\right] \Longrightarrow \tau_{\text{diff}} = \mu_0 \sigma L^2, \tag{6}$$

where L is a typical scale of length.

In the case where conductivity is very large, or movements are very fast, the diffusive term can be neglected,

$$\dot{\boldsymbol{B}} = \boldsymbol{\nabla} \times \boldsymbol{u} \times \boldsymbol{B}. \tag{7}$$

As we will see in the following few paragraphs, this equation implies that the magnetic flux through the surface S spanned by a closed curve C comoving with the plasma is conserved. Figure 1 shows a sketch of the situation. The magnetic flux passing through S can change in two ways. On the one hand, the field strength $\boldsymbol{B}$ within the curve C can change. On the other hand, the curve C can move with respect to the field $\boldsymbol{B}$. In the first possibility, the change of flux in a surface element $d\boldsymbol{A}$ is given by

$$\dot{\boldsymbol{B}} \cdot d\boldsymbol{A} \tag{8}$$

and the total change is just the integral over the entire surface S. In the second possibility an infinitesimal line element $d\boldsymbol{l}$ of the curve moves relative to $\boldsymbol{B}$. The change in the enclosed field is then given by

$$\boldsymbol{B} \cdot (\boldsymbol{u} \times d\boldsymbol{l}) \tag{9}$$

and the total change is the contour integral along C. Using the vector identity $\boldsymbol{A} \cdot (\boldsymbol{B} \times \boldsymbol{C}) = (\boldsymbol{A} \times \boldsymbol{B}) \cdot \boldsymbol{C}$ we can rewrite (9) and write the entire change in the flux through the surface S spanned by C as

$$\frac{d}{dt} \int_S d\boldsymbol{A} \cdot \boldsymbol{B} = \int_S d\boldsymbol{A} \cdot \frac{\partial \boldsymbol{B}}{\partial t} - \oint_C d\boldsymbol{l} \cdot (\boldsymbol{u} \times \boldsymbol{B}) \,. \tag{10}$$

We rewrite the contour integral as a surface integral using Stokes' theorem

$$\oint_C d\boldsymbol{l} \cdot (\boldsymbol{u} \times \boldsymbol{B}) = \int_S d\boldsymbol{A} \boldsymbol{\nabla} \times (\boldsymbol{u} \times \boldsymbol{B}) \,. \tag{11}$$

Now the integrand is exactly the right-hand side of the induction equation for the case of infinite conductivity. This implies that

$$\frac{d}{dt} \int_S d\boldsymbol{A} \cdot \boldsymbol{B} = \int_S d\boldsymbol{A} \cdot \left(\frac{\partial \boldsymbol{B}}{\partial t} - \boldsymbol{\nabla} \times (\boldsymbol{u} \times \boldsymbol{B}) \right) = 0, \tag{12}$$

proving that the flux through the curve C remains unchanged. The physical meaning of this is that the magnetic field co-moves with the plasma, this is often called the "frozen-in magnetic field" , or ideal MHD. The transition from a diffusion dominated to a frozen in situation can be parameterized by the ratio of diffusion time τ_d to convection time τ_u, where

$$\tau_d = \mu_0 \sigma L^2, \qquad \text{and} \qquad \tau_u = \frac{L}{u}. \tag{13}$$

We can now define the magnetic Reynolds number R_M,

$$R_M = \frac{\tau_d}{\tau_u} = \frac{\mu_0 \sigma L^2 u}{L} = \mu_0 \sigma L u. \tag{14}$$

For large R_M the field is frozen in, for small R_M it diffuses. In general, R_M is large in the interplanetary medium, as well as in the chromosphere and corona.

The lesson from this paragraph is that magnetic field can diffuse, if conditions are appropriate. From what we have learned so far, field lines will remain connected ("frozen in") if conductivity is large enough, when diffusion dominates, we can envisage field lines swapping their identity in localized regions of low conductivity or where R_M is not much larger than unity. This can happen, when field is brought into a region where it needs to change on a very small scale, e. g. near a current sheet. Then the diffusion time can be short compared to the convection time and the "frozen in" approximation breaks down. A field line can then disconnect from its original footpoint and then reconnect with another field line, as sketched in Fig. 2. While the "frozen in" approximation is still valid in the global context, it can be broken locally which allows for changes in magnetic topology, as is shown in Fig. 2. To summarize, reconnection can occur when there is a global configuration that brings magnetic field together at a rate and on a scale sufficient

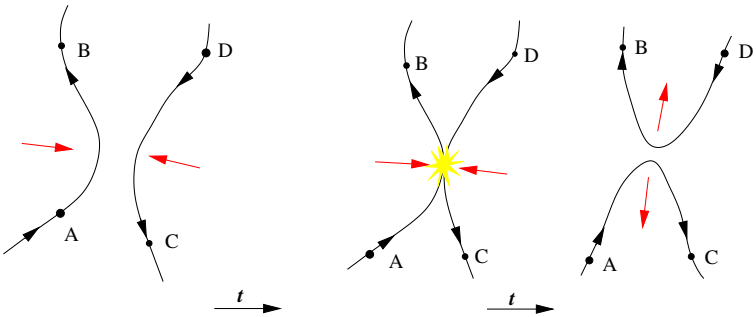


Fig. 2. Reconnection separates fluid elements that were originally connected and connects them with new ones. This can lead to changes in magnetic topology. In such regions (sketched with a flash here) scale lengths are short, e. g. because the polarity of the magnetic field changes on a small scale.

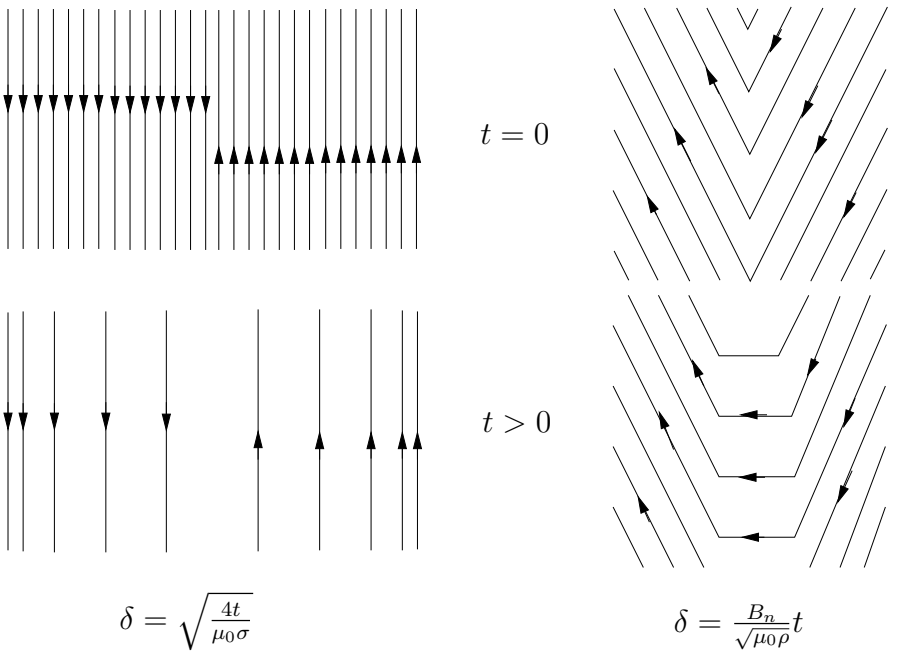


Fig. 3. Reconnection in the model of Parker and Sweet (left) and Petschek (right).

for breakdown of the "frozen in" approximation. We will now consider the original model of Parker and Sweet, consider its difficulties and improve the situation by introducing Petscheks model for reconnection. A very primitive sketch of the two models is shown in Fig. 3.

2.2 The Model of Parker and Sweet

Consider a situation as sketched in Fig. 3a where magnetic field of opposite polarity can be transformed into energy by diffusive annihilation. The corresponding diffusion equation

$$\frac{\partial \boldsymbol{B}}{\partial t} = \frac{1}{\mu_0 \sigma} \Delta \boldsymbol{B},$$

has the one-dimensional solution

$$B = \sqrt{\frac{\mu_0 \sigma}{4t}} e^{-\frac{y^2 \mu_0 \sigma}{4t}}. \tag{15}$$

The field diffuses and locations of field strength $1/e$ move away from the current sheet at a speed $v_{1/e} = \sqrt{4/\mu_0 \sigma t}$, as sketched in Fig. 4. The field strength in the current sheet diminishes, the difference to the original field strength going into the energy liberated by reconnection. This can be a substantial amount of energy, visualized here by a dramatic change in field strength. However, we can do better than that, and derive an estimate of the reconnection rate, the speed at which magnetic field can be brought into the reconnection region. For this, we consider the very much simplified geometry sketched in Fig. 5. Plasma flows in from the top and the bottom at a speed u_{y_0} over a width $2L$ and needs to exit the reconnection region (the gray shaded area in Fig. 5) of width 2δ at a speed v which can be determined from the equation of continuity,

$$u_{y_0} L = v\delta, \tag{16}$$

This speed v can also be estimated from the Bernoulli equation

$$\frac{\rho v^2}{2} = p - p_0, \tag{17}$$

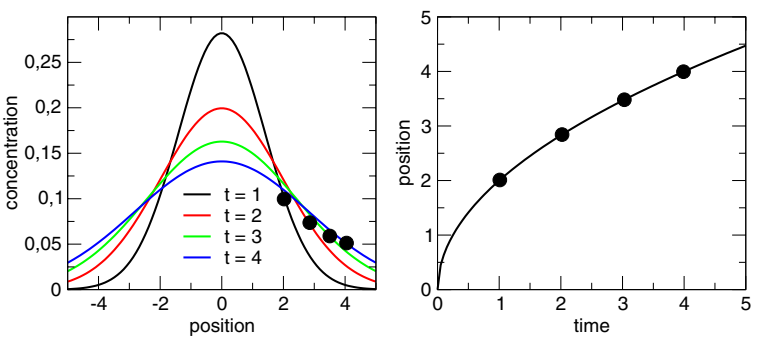


Fig. 4. Diffusion: The left-hand panel shows four Gaussians at successive time steps ($t = 1$ through $t = 4$) with big dots at $1/e$. The position of the dots is plotted versus time in the right-hand panel, illustrating the meaning of diffusion speed.

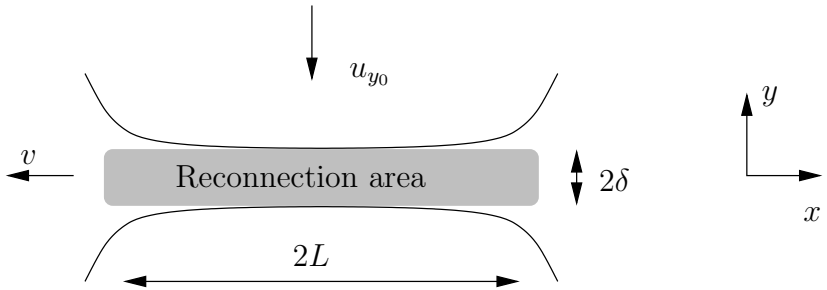


Fig. 5. Geometry of reconnections and the equation of continuity.

where ρ is the plasma density, p the pressure in the center of the reconnection region, and p_0 the pressure in the undisturbed region far away. Of course, we're neglecting magnetic pressure here, in general this is of the same magnitude as the plasma pressure ($\beta \approx 1$), so we're introducing an error of the order of a factor of two. It too, will push the plasma outwards along the current sheet. In order for the reconnection region to remain in hydrostatic equilibrium with the surrounding plasma, the pressure at the boundary needs to be determined by

$$p - p_0 = \frac{B_{x_0}^2}{2\mu_0}, \tag{18}$$

where B_{x_0} is the field strength in the undisturbed region.

Next, we derive an expression relating the inflow speed u_{y_0} to the diffusion speed in the reconnection region. We compute the time is takes diffusion to traverse a layer of thickness δ. From

$$\delta = \int_0^T \mathrm{d}t \sqrt{\frac{1}{4\mu_0 \sigma t}}$$

we obtain the diffusion speed v_{diff}

$$v_{\mathrm{diff}} = \frac{\delta}{T} = \frac{1}{\mu_0 \sigma \delta}. \tag{19}$$

The inflow speed u_{y_0} can be estimated in the following way. Inside the reconnection region there must be an X-point because of symmetry considerations. There, B, as well as flow speeds must be very small and hence Ohms law can be written as

$$J_z = \sigma E_z,$$

where J_z is the current density and E_z the electric field. Because σ is a scalar, E_z needs to be along J_z and point along the z-axis. For a stationary flow $\dot{B} = 0$ implies than $\nabla \times E = 0$ and hence E_z is constant. Hence the field can be estimated using quantities valid outside the reconnection region, where Ohms law tells us that

$$E_z = -u_{y_0} B_{x_0}.$$

Because the entire change of the magnetic field over the reconnection region must amount to $2B_{x_0}$, the magnitude of the total current carried by the current sheet can be estimated

$$\mu_0 J_Z = \frac{2B_{x_0}}{2\delta}.$$

Combining the last three equations, we obtain an estimate for the inflow speed u_{y_0},

$$u_{y_0} = \frac{1}{\mu_0 \sigma \delta}, \tag{20}$$

i. e. the inflow speed is just the diffusion speed! All we need to do now to estimate whether this is fast or slow, is to compare this with the outflow speed v and relate that to the overall geometry. Inserting (18) in (17) we find that it is just the Alfvén speed, $v_A = B_{x_0}/\sqrt{\mu_0 \rho}$, (as we could easily have guessed). We estimate the unknown width of the reconnection region, δ, using our knowledge of the diffusion speed (20)

$$\delta = \frac{1}{\mu_0 \sigma u_{y_0}},$$

which we insert in the continuity equation, (16), and solve for the inflow speed u_{y_0}

$$u_{y_0} = \left(\frac{B_{x_0}}{\sqrt{\mu_0 \rho}} \frac{1}{\mu_0 \sigma L} \right)^{1/2}. \tag{21}$$

This can also be written as

$$u_{y_0} = v_A \left(\frac{1}{v_A} \frac{1}{\mu_0 \sigma L} \right)^{1/2}, \tag{22}$$

where the expression in parenthesis is just the inverse of the magnetic Reynolds number, and hence

$$u_{y_0} = v_A \, R_M^{-1/2}.$$

In other words, reconnection is a slow process in the Parker and Sweet configuration because the magnetic Reynolds number is a large number in general.

2.3 The Petschek Model

Comparison of light curves of flares on the Sun and on other astrophysical objects with the time scale expected from a Parker and Sweet-like reconnection configuration showed that this is too slow by at least two orders of magnitude. Petschek (1964) [36] solved this difficulty by allowing for standing Alfvén waves which deflect the magnetic field far away from the actual reconnection region, as sketched in Figs. 3 and 6. These waves allow for a

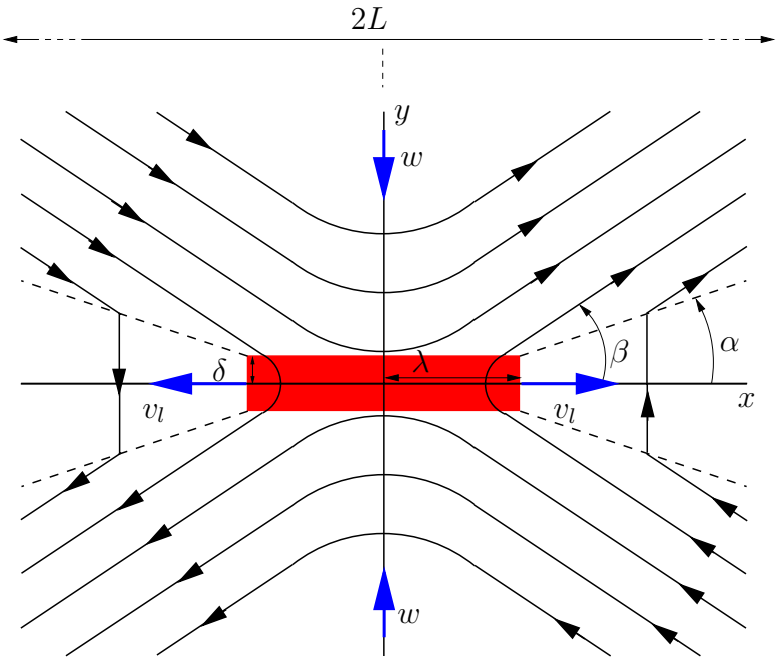


Fig. 6. The geometry of reconnection according to Petschek. Reconnection takes place in the rectangle $2\lambda \times 2\delta$. The field is transported into the reconnection area with a speed w and over a width $2L$. This leads to a formation of standing Alfvén waves (dashed lines) at an angle α to the x axis. The field is parallel to the x axis far away from the reconnection area but is bent to an angle β with respect to the x axis in the vicinity of the reconnection area.

much smaller reconnection region than the width of the inflow region, resulting in a much higher inflow velocity. We denote by 2λ the total width of the reconnection region, and by 2δ its total thickness. Inside the reconnection region, continuity requires

$$w = v_A \frac{\delta}{\lambda},$$

however, the situation is quite different outside the reconnection region proper, where the standing Alfvén waves deflect the inflowing plasma. There continuity only requires

$$w = v_A \tan \alpha$$

which can be a substantial fraction of the Alfvén speed, depending on the unspecified angle α. Continuity of the flow across the boundary of the reconnection region requires that

$$\frac{w}{v_A} = \frac{\delta}{\lambda} = \tan \alpha.$$

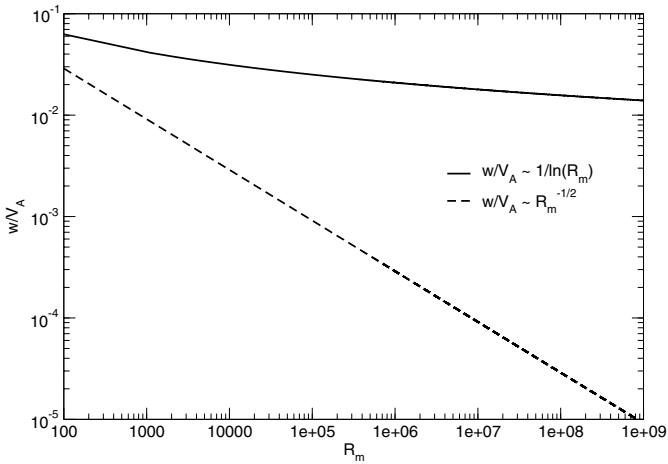


Fig. 7. Comparison of the reconnection speeds in the model of Petschek (solid line) and the model of Parker and Sweet (dashed line).

This angle α can't be very large, because otherwise the tension force (determined largely by the angle β in Fig. 6) on the inflowing magnetic field would prevent it from reaching the reconnection region. Hence there must be an optimal value for λ at which the inflow speed is maximized (but still smaller than v_A). Petschek found a way of estimating this λ as a combination of the angles α and β, as well as of other plasma quantities. As the derivation is somewhat lengthy and involved, we only give the final result for the inflow speed, or reconnection speed, which is possible in a Petschek-type configuration,

$$\frac{w}{v_A} \approx \frac{\pi}{4\,e\,\ln R_m}. \tag{23}$$

The reconnection speed only decreases with the logarithm of the magnetic Reynolds number. As is illustrated in Fig. 7, this makes a great difference when we consider astrophysical plasmas, for which R_M is generally large.

The reconnected field lines are not only expelled from the reconnection region by the outflowing plasma, they also feel a tension force acting on them which is due to the Lorentz force

$$\boldsymbol{F} = q\,(\boldsymbol{E} + \boldsymbol{v} \times \boldsymbol{B})\,. \tag{24}$$

acting on the particles constituting the plasma. This microscopic force is better replaced by a macroscopic quantity, the force $\mathcal{F}$ acting on a unit volume of plasma. This can be written as

$$\mathcal{F} = \nabla \cdot \mathcal{M} - \epsilon_0 \frac{\partial(\boldsymbol{E} \times \boldsymbol{B})}{\partial t}, \tag{25}$$

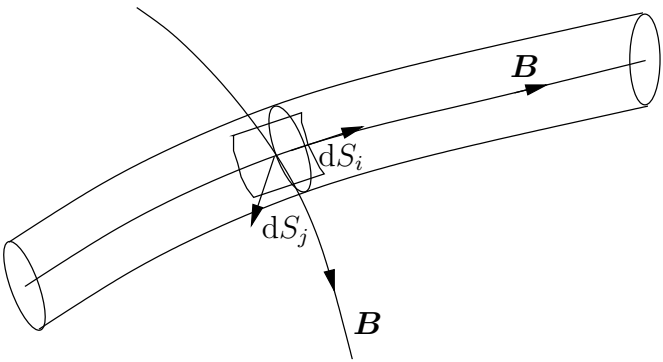


Fig. 8. The Maxwell stress tensor $\mathcal{M}_{ij}$ describes the tension which the electromagnetic field in the j direction exerts on the field in the i direction.

where the second term is just the temporal change in the Poynting flux and $\mathcal{M}_{ij}$ is the Maxwell stress tensor,

$$\mathcal{M}_{ij} = -\delta_{ij}\left(\frac{\epsilon_0}{2}E^2 + \frac{1}{2\mu_0}B^2\right) + \epsilon_0 E_i E_j + \frac{1}{\mu_0}B_i B_j \tag{26}$$

The first term describes the isotropic pressure exerted on the plasma by the magnetic field. Because the electric field vanishes under normal circumstances in the heliosphere, we neglect it in the following considerations of the remaining part of $\mathcal{M}$, $1/\mu_0\, B_i B_j$. It describes a tension. Consider a small surface $\mathrm{d}\boldsymbol{S}$ in Fig. 8 whose normal vector points along the tangent to the magnetic field. Obviously, the i component of the surface normal $\boldsymbol{n}$ is then $n_i = B_i/B$. The tension acting on $\mathrm{d}S_i$ (and hence along B_i) is

$$\mathrm{d}S\, n_j\, B_i B_j/\mu_0 = \mathrm{d}S\, B_j\, B_i\, B_j/(\mu_0 B) = n_i\, \mathrm{d}S\, B^2/\mu_0. \tag{27}$$

The sign in the tension term is opposite of that in the pressure term which shows us that the two are indeed something else. $\mathcal{M}_{ij}$ is the force per unit area in the i direction exerted by the j component of the field. The force on the field within the flux tube depicted in Fig. 8 is exerted on it by the j component of the field coming from above $\boldsymbol{B}$ over the positive side of the surface $\mathrm{d}S_j$ onto the field on the negative side of the surface (inside the flux tube). Hence we may envisage bent field lines as rubber bands under tension. When the holding force disappears (as it does after the field lines leave the diffusion dominated reconnection region), they try to straighten out thus acquiring an energetically more favorable state. Reconnected field lines will thus tend to move away from the reconnection region and pull the plasma with them.

The process of reconnection, of outflowing plasma and contracting field lines probably generates a very turbulent medium in the vicinity of the reconnection site. It will tend to generate this on a large spatial scale (small

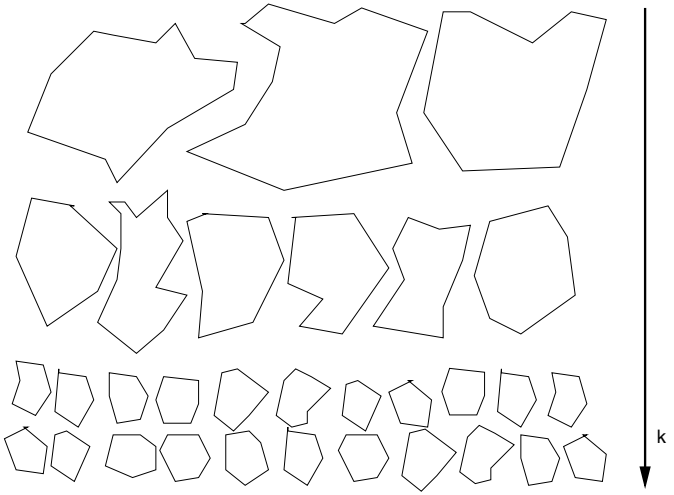


Fig. 9. Turbulent cascade. Large-scale structures disintegrate into smaller structures accompanied by an increase in the wave vector $\boldsymbol{k}$. The large structures contain the most power.

wave number k) and the enormous amount of turbulent energy will cascade in a Kolmogorov-type cascade to higher and higher wave numbers k where it is absorbed by ions with a low Q/A ratio. The energy that is not absorbed is available to ions with a higher Q/A. Figure 9 shows a cartoon of the cascade process.

2.4 The Ejection of Mass

Cool loops (prominences) are visible in emission lines of neutral hydrogen against the solar limb. Against the solar disk they appear in absorption as so-called filaments. Typically, these entities are about 5'000 km thick, 50'000 km high, and 200'000 km long. Because they are about 100 times cooler than the material in the corona into which they penetrate, their density needs to be about 100 times higher than that of the surrounding corona to allow pressure balance. They often remain unchanged for weeks to months and hence we may assume static equilibrium for simplification. Because of their low temperature, this equilibrium cannot be simply hydrostatic because the scale height is much too small. However, filaments can be held in equilibrium by electromagnetic forces. We know from photospheric magnetograms that they appear to track lines along $B_r = 0$ and hence there are three distinct topologies of the magnetic field which could support the filament. They are illustrated in Fig. 10. In the simplest configuration (left) the tension of the field lines due to the filament counteracts gravitation. The field needed to supply the Lorentz force is rather small. Following Golub and Pasachoff (1997) [16]

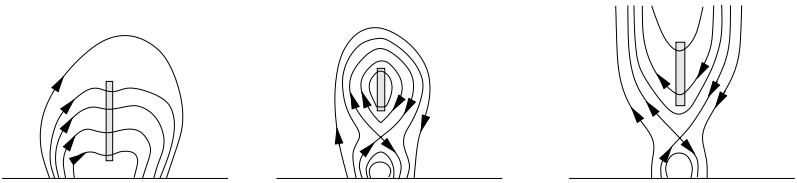


Fig. 10. Three different magnetic topologies that can support a prominence/filament.

we consider the prominence as a thin sheet which extends to infinity along the y axis. In magnetohydrostatic equilibrium we have

$$-\frac{\mathrm{d}P}{\mathrm{d}z} - \rho g + (\boldsymbol{J} \times \boldsymbol{B})_z = 0 \tag{28}$$

in the z direction, i. e. the vertical. We assume there are no forces acting in the other directions. The pressure gradient is negligible, the typical scale height is about 300 km and hence this does not contribute a significant amount to the total force. We integrate the remaining terms across the prominence (in the x direction, to the right in Fig. 10).

$$g \int \mathrm{d}x \rho = \int \mathrm{d}x (\boldsymbol{J} \times \boldsymbol{B})_z. \tag{29}$$

We can replace $\boldsymbol{J}$ by $1/\mu(\nabla \times \boldsymbol{B})$ using Ampéres law and hence only the x and y components of $1/\mu(\nabla \times \boldsymbol{B})$ enter the integrand.

$$\begin{aligned} g \int \mathrm{d}x \rho &= \frac{1}{\mu} \int \mathrm{d}x \left((\nabla \times \boldsymbol{B}) \times \boldsymbol{B} \right)_z, \\ &= \frac{1}{\mu} \int \mathrm{d}x \left((\frac{\partial B_z}{\partial y} - \frac{\partial B_y}{\partial z}) B_y - (\frac{\partial B_x}{\partial z} - \frac{\partial B_z}{\partial x}) B_x \right), \\ &= \frac{1}{\mu} \int \mathrm{d}x B_x \frac{\partial B_z}{\partial x}, \end{aligned} \tag{30}$$

because the other terms vanish. B_x is constant across the narrow sheet and can be put in front of the integral, leaving only the difference in the z component of $\boldsymbol{B}$,

$$g \int \mathrm{d}x \rho \approx \frac{1}{\mu} B_x \left[B_z \right], \tag{31}$$

where $[B_z]$ is the jump in the z component of $\boldsymbol{B}$ across the prominence. This jump does not need to be large; for $\Delta x = 5000$ km, $\rho = 10^{-10}$ kg/m^3 and $B_x = 10^{-3}$ T we have $[B_z] \approx 2 \cdot 10^{-4}$ T, which is indeed small, even so small, that it is not easily measured.

Even the direction of the x component of the field is hard to measure. Using statistical inferences, one finds that the configuration sketched in Fig. 10a

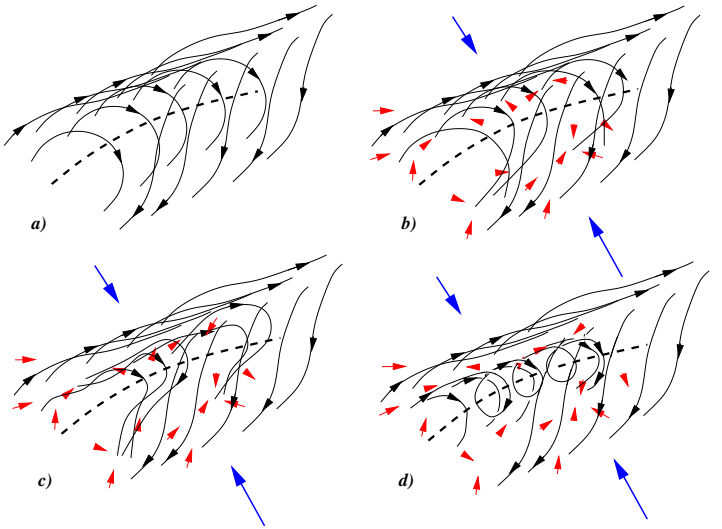


Fig. 11. The motions in the granulation (short arrows) and a systematic motion towards the neutral line (long arrows) can lead to the formation of a flux rope.

does not describe the situation on the Sun, but that the x component of the field points in the other direction, as illustrated in Fig. 10b and c. This implies that there is a point (or a long line extending along the y direction) where the field "crosses", an X-point. This naturally implies reconnection at some point.

Figure 11 illustrates how a configuration like Fig. 10b or c can be achieved. The motion of the footpoints of the field lines together with a systematic motion towards the neural line can lead to the formation of a flux rope through magnetic reconnection. Such a configuration can explain the three-part structure observed in CMEs that we will encounter in the next section. A dense shell leads the CME, followed by a cavity which contains the prominence at the trailing end. In the picture discussed here, the cavity is a consequence of the long connecting path to the photosphere along field lines. Because of their length prominences are well isolated from the mass and energy supplying photosphere. The energy in this high-field region is insufficient to heat the plasma to a temperature corresponding to a scale height sufficient to fill the entire void or cavity. In the surrounding magnetically open regions the plasma is hot, the scale height is large and these regions are thus filled with plasma. Inside the cavity, the material "condenses" towards the bottom and forms the prominence. The total mass can be large, up to the entire mass of the originally filled flux rope (what is now the cavity). The gravitative force on this mass holds the flux rope down until ongoing footpoint motions have increased the field strength to the point where magnetic buoyancy exceeds gravitation and the flux rope can erupt to form a CME. This scenario is only

one of many possible scenarios explaining the onset of CMEs, I like it because it is intuitively appealing. However, it is not without problems, why should the granulation motion conspire to move the footpoints of the field lines in such a way that they will form a flux rope? Why does CME frequency correlate so well with solar activity (see Fig. 21)? Low (1997) [27] has introduced a promising scenario that closely couples CME initiation with the emergence of magnetic flux from the new magnetic cycle. The emerging new flux has the opposite polarity and will then reconnect with the field of the old polarity, leading to the formation of CMEs. In fact, Nindos and Zhang (2002) [32] and Nindos et al. (2003) [33] found that observed photospheric shearing motions (even when they are large) inject far less helicity than is removed by a CME. Consequently, they state, the main source of the helicity carried away by the CME is the new magnetic flux that emerges twisted from the convection zone. Hence shearing in the photosphere is not the mechanism that builds up the magnetic free energy that will ultimately lead to CME initiation.

3 Coronal Mass Ejections

3.1 (Mostly) Remote Observations

White-light observations are the classical observations that have allowed studies of CMEs since the Solar Maximum Mission (SMM). Photospheric light is Thomson scattered by coronal electrons. The scattered light is most intensive in regions of high density, and hence images in white light are especially useful to derive the coronal density and changes in it. The often-cited three-part structure of CMEs dates back to SMM observations and the name "coronal mass ejection" originated from white-light images. The original definition was a mass ejection from the Sun observed in white light. From several sequential images one can derive e. g. velocity curves of CMEs. A typical sequence of SMM observations is shown in Fig. 12.

These ejections are accompanied by some of the most spectacular phenomena in the solar system. Remarkably, their average speed close to the Sun is slower than that of the solar wind at 1 AU. It ranges from about 270 km/s around solar activity minimum to about 470 km/s around solar activity maximum. The speed distributions have a distinct tail towards higher speeds which can exceed 1000 km/s. On average, several times 10^{12} kg of material are ejected in a CME (this corresponds to a cube of water with a side length of more than 1 km). The corresponding kinetic energy is a few times 10^{23} J, dividing by a typical duration on the order of an hour gives us the power, roughly 10^{20} W, which corresponds to the electric power output by about 10^{11} nuclear power plants. Regardless of this large number, this output pales against the total power output of the Sun, $S_0 \times 4\pi\,(1.5\,10^{11})^2 \approx 4 \times 10^{26}$ W.

After the swelling of the original streamer, the earliest sign of an imminent eruption is the disappearance of the filament ("disparition brusque"), announced by movements indicating that the magnetic configuration of the

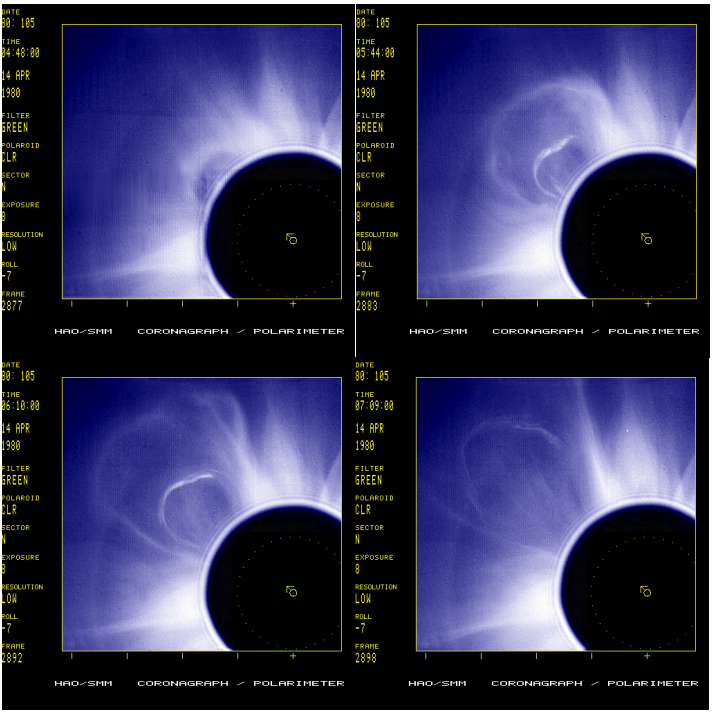


Fig. 12. SMM White-light images of a prototypical coronal mass ejection observed on April 14, 1980. The outermost loop is moving away from the Sun at a plane-of-sky speed of 409 km/s, the prominence at 365 km/s, and the cavity beneath the prominence at 353 km/s.

prominence is becoming unstable. Often the filament erupts as an arch, speeds up to several hundreds of km/s have been observed. Against the limb the impressive prominences look very spectacular. The field lines beneath the expanding prominence rapidly reconfigure into a system of arcs that are oriented more or less perpendicularly to the neutral line. The forming system of arcades is very hot and emits in extreme ultraviolet and in X-rays.

The temporal sequence of events associated with CME onset, initiation, and evolution is a subject of intensive debate. Observationally, the matter is much simpler. Figure 13 gives an overview of the types of radiation associated with the eruptive phenomena observed at the Sun. In the following, we will discuss these observations keeping in mind the very much idealized situation illustrated in Fig. 14.

- Radio waves can be measured with receivers on Earth in certain frequency bands, however, for the part of the spectrum that is reflected by the ionosphere (and that is important for terrestrial radio communications) space-based receivers are needed. Depending on the base line of

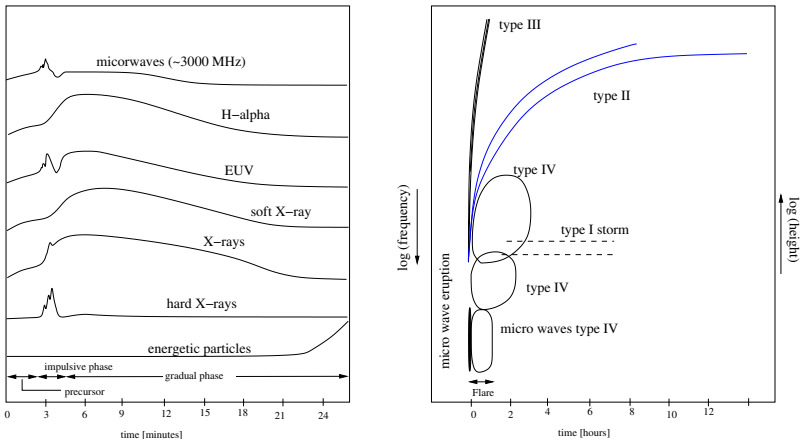


Fig. 13. Temporal sequence of electromagnetic and particle emission associated with the eruption of a filament, accompanied by a flare. After Dulk (1985) [10]. Note the very different time scales in the two panels.

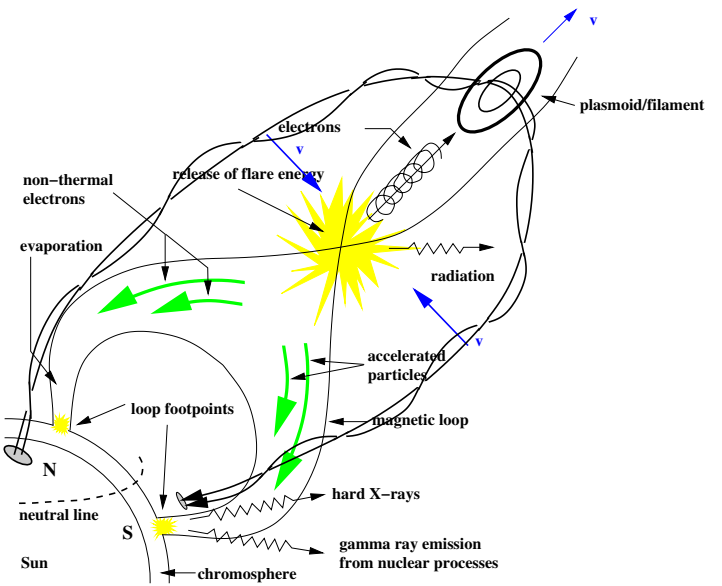


Fig. 14. Model of a flare on the Sun, after Lang (2000) [23] with modifications.

the receivers, one can even obtain spatially resolved images of the Sun or of active regions. In general, space-based receivers don't do this, however, combinations of various instruments on different spacecraft can be used to measure the interplanetary evolution of CMEs. Various different types of radio emission are distinguished:

  - Type I: Long-lasting sources of radio emission (hours to days) with temperatures up to 10^6 to 10^7 K. These radio waves are generally not emitted by CMEs or flares, but by the acceleration of low-energy electrons in large magnetic loops which connect active regions with more distant regions on the Sun.
  - Type II: Eruptions of metric radio emission (frequencies between 0.1 and 100 MHz). This emission begins at high frequencies and slowly drifts to lower frequencies. This emission is interpreted as the emission of the plasma at the plasma frequency ($\sim 9000\sqrt{n_p}$ for n in cm^{-3}). Often the second harmonic is also measured. The slow drift corresponds to speeds through the corona of about 1000 km/s which is about the speed of CMEs and hence the emission is believed to be due to a shock wave that is driven through the corona by the CME. The emission can be detected at Earth because density is lower everywhere in between the emitting site and Earth and hence there is no absorption between the emitting site and Earth.
  - Type III: This kind of metric radio emission (frequencies between 0.1 and 1000 MHz) is nearly always seen accompanying flares. In contrast to type II eruptions, the frequency shift is much more rapid (up to 100 MHz per second). The emission presumably stems from electrons which have been accelerated away from the Sun with energies between 10 and 100 keV (half the speed of light). Type III emission is generated by the synchrotron motion of electrons when they spiral along the magnetic field. This radiation is probably at least partially converted into other radio waves. Nevertheless, the radiation is well focused along the direction of electron propagation at these energies, and hence the radiation is more or less perpendicular to the magnetic field and acts much like the light beam of a light house.
  - Type IV: Wide banded continuum radiation which can last up to several hours after a flare. It is presumably due to energetic electrons that are confined in a magnetically closed system. They emit synchrotron radiation whose source is transported into the interplanetary medium by the magnetic cloud.
  - Type V: Appears as an eruption after a type III eruption, but does not exhibit a frequency shift. The source probably lies in the coronal plasma itself.
- Hα: The disappearance or eruption of the filament is often accompanied by an increase in the intensity of H-α. The H-α line of neutral hydrogen (a substantial fraction of the the filament consists of neutral hydrogen) is a chromospheric line, the increase is often called chromospheric brightening. It was first discovered in 1859 (!) by R. C. Carrington and R. Hodgson in a region with a complex sunspot configuration. This chromospheric brightening or flare appears in the form of two ribbons parallel to the neutral line which move apart at a speed of about 10 km/s, corresponding to loops at increasing heights or reflecting reconnection of increasingly

distant field lines. The space between the ribbons is filled by the arcade structure (in Hα as well as in other wave lengths). The chromospheric flare lasts a few minutes and the emission decays in the course of an hour. Some flares can even be seen in white light.

- EUV: The strong heating of the plasma in a reconnection region leads to the emission of transition lines of highly ionized elements. Because the transitions are very energetic, the corresponding wavelength is short and lies in the extreme ultraviolet (EUV). A typical ion is the He-like Fe XVII which emits at 1.5 nm, corresponding to photons with an energy of about 1 keV. Both EIT on SOHO and TRACE can obtain high resolution images at this wavelength at high cadence, allowing us to observe the temporal evolution in some detail.
- soft X-rays: The high temperature of the plasma (up to 10^7 K) is reflected in the temperature of the electrons. This corresponds to electron thermal speeds up to 5% of the speed of light. When these electrons collide with protons or other ions, they are strongly deflected, i. e. accelerated and emit thermal Bremsstrahlung. This lies in the region of soft X-rays. Soft X-ray images by SMM and SXT on Yokoh were the first to be used to observationally identify regions of magnetic reconnection.
- Hard X-rays: Radio observations with the Very Large Array facility had already shown in the eighties that energetic electrons are generated in the vicinity of the apex of coronal loops. The often observed frequency shift towards higher frequencies was interpreted as due to electrons that were accelerated downwards into the corona. HXT on Yokoh confirmed this notion in an impressive way as well as the the existence of two additional sources at the footpoints of the loops. 50% of all cases observed by HXT showed such a double source, the other half was divided into small, possibly unresolved sources or very large sources which had erupted simultaneously, possibly leading to more complex magnetic configurations. The hard X-rays are due to Bremsstrahlung from electrons that have been accelerated to high energies. The energy of the radiation is too high to be thermal, if so, temperatures exceeding 10^9 K would be required.
- γ-radiation: The energetic nature of eruptions on the Sun even allows nuclear reactions. Currently, RHESSI is a source of exciting new findings about the nature of the high-energy end of these energetic phenomena. That nuclear reactions must take place in the solar atmosphere had been known for some time, for instance because of the observation of the electron-positron annihilation line at 511 keV. Lines at higher energies have also been observed, some very light elements such as Be and Li can be produced,

$$^4\mathrm{He} + \alpha \rightarrow^7 \mathrm{Be} + n + \gamma\,(0.431\mathrm{MeV}),$$

or

$$^4\mathrm{He} + \alpha \rightarrow^7 \mathrm{Li} + p + \gamma\,(0.478\mathrm{MeV}),$$

where the excited state decays and emits a γ. Nuclear reactions also lead to the creation of neutrons which were first observed by the neutron monitor on the Jungfraujoch (Debrunner et al., 1983 [9]). High energy protons ($E > 300$MeV) interact with the solar atmosphere to produce mesons which in turn decay accompanied by emission of γ radiation. The temporal profile of hard X-ray and γ radiation are nearly the same, indicating that electrons and ions are accelerated at the same time, although not necessarily at the same location, as has been shown by RHESSI (Hurford et al., Lin et al., 2003, 2003 [19, 25]).

- energetic particles: Reconnection accelerates particles which can either interact with the solar atmosphere or escape to interplanetary space. The expansion of the magnetic field leads to a focusing of the particles because the first adiabatic invariant, the magnetic moment, is conserved. Thus they escape from the Sun with a small pitch angle and stream along the magnetic field. These accelerated particles are thus confined to relatively narrow flux tubes in the heliosphere. Often the electrons reach the Earth with only a very short delay. They travel along a path that is only little longer than the path along the ideal archimedan spiral described by the interplanetary magnetic field. The path can be slightly longer due to their finite pitch angle. The same holds for energetic protons, however, at equal energy, the electrons will of course arrive much sooner. Electrons can be accelerated up to energies of about 100 MeV, protons up to 1 GeV or higher.

  Obviously, particles with a high energy must arrive at an observer before lower energy particles. This leads to velocity dispersion that can be used to derive the path length the particles traveled, or assuming a path length, to determine when the particles were accelerated. As a rule of thumb, one assumes a path length of 1.2 AU, corresponding to typical solar wind speeds. Interestingly, path lengths up to 2 AU have been observed, implying connection to a site beyond the Earths orbit or a strong deformation of the magnetic field between the Sun and the Earth e. g. due to a CME. Because electrons have a much larger mean free path than protons at the same energy, they are excellent tracers of the magnetic configuration of the heliosphere.

3.2 Solar-Cycle Dependence

The solar cycle has been known for more than a century and a half (Schwabe, 1843 [39]; Wolf, 1856 [62]). The word cycle is used to underline that solar activity is not periodic, the average duration of 11 (or 22) years of solar activity is an average and does exhibit certain variability. Probably because of their obviousness and the availability of long observational time series, sunspots have long been mentioned in one with solar cycle variations. Figure 15 shows a reconstruction of yearly sunspot numbers dating back to nearly 1600. The apparent prolonged minimum in sunspot number lasting from about 1645 to

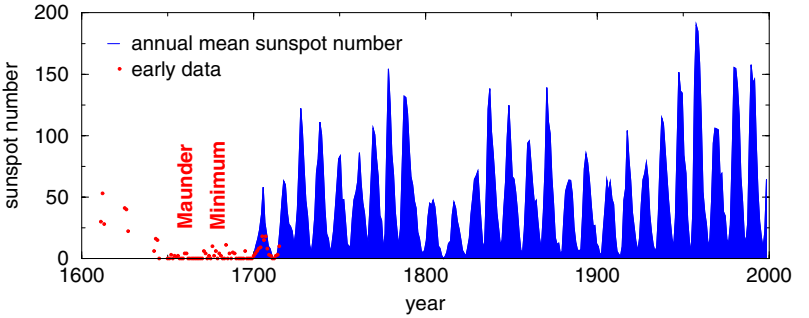


Fig. 15. Yearly averaged sunspot numbers back to 1600, clearly showing the ∼11 year solar cycle. Visible is the ∼ 11 year solar cycle and the Maunder minimum around 1645 - 1700.

about 1720 is real. It has been termed the Maunder minimum (Spörer, 1887 [47]; Spörer, 1889 [48]; Maunder, 1890 [28]; Maunder, 1894 [29]; Edddy, 1976 [11]. During such sunspot minima the regular solar activity cycle of about 11 years all but disappears as measured by the yearly averages of sunspot numbers. However, other phenomena such as the modulation of galactic cosmic rays persist, even during the Maunder minimum (Beer et al., 1998 [4]).

The solar activity cycle is not limited to a variation in sunspot number. Sunspot number (and thus solar activity) is closely linked with the solar 10.7 cm radio flux, the number of Hα flares, metric type II radio bursts, disappearing filaments, or erupting prominences. Turning to interplanetary space, the structure of the inner heliosphere is strongly influenced by solar activity. This manifestation of the solar dynamo influences the interplanetary magnetic field, the types of solar wind flowing through the heliosphere, modulates the galactic cosmic rays and the anomalous cosmic rays. Also, it largely determines the number of interplanetary shocks at 1 AU (Lindsay et al., 1994 [26]; Smith, 1983 [45]) (see Fig. 22). While the structure of the heliosphere is relatively simple during solar activity minima, the situation is much more complex during activity maxima. During solar activity minimum, slow and fast solar wind streams interact and form so-called corotating interaction regions which largely determine the large-scale structruee of the heliosphere (Balogh et al., 1999 [3]). High-speed streams emanate from the polar coronal holes while the source of the slow wind is concentrated along the streamer belt. In the expansion into interplanetary space, stream-stream interactions set up corotating interaction regions which develop strong shocks beyond about 2 AU and are prolific particle accelerators. During solar activity maximum, matters are much more complicated because the solar corona undergoes violent changes while restructuting itself to its new magnetic polarity. The most spectacular manifestations are certainly the coronal mass ejections which appear carry with them the heliospheric current sheet (Crooker et al.,

1998 [8]). These are emitted a few times a day, and the shocks driven by them, when present, can accelerate particles to high energies, so-called SEPs from gradual events.

Of course, we do not have a large database of interplanetary observations dating back many activity cycles, nevertheless, the physical understanding of many of the processes linking heliospheric behavior with solar activity have been established well enough that we can compare present-day observations with conditions in the past. Solar activity may be an important factor in understanding the "SEP" population in lunar soils (Wieler et al., 1986 [54]; Wieler, 1998 [53]). It has often been attributed to solar flare material which, in more modern terminology, would probably be called SEPs from impulsive events, although other interpretations have been advanced (Wimmer-Schweingruber and Bochsler, 2001 [60]).

3.3 Compositional Aspects

Abundance measurements in the solar wind, energetic particles, but also remote coronal observations with spectrometers such as SUMER or UVCS on SOHO (Wilhelm et al., 1995 [55]; Kohl et al., 1995 [22]) show a pronounced and systematic difference in abundances when compared to photospheric or solar abundances. Elements with a low first ionization potential (FIP $\lesssim 10$) are enhanced by a factor between 3 and 5 with respect to elements with a high FIP (> 11.5 eV)[1]. Active regions on the Sun appear to behave somewhat differently, FIP biases (the abundance ratio of low-FIP elements relative to high-FIP elements) as high as 16 have been observed, depending on the height of the loops investigated. An emerging active region appears to have more or less photospheric abundances which are then soon modified to reach values more typical of the corona within a day or two. Widing and Feldman (2001) [52] used Skylab spectroheliograms to determine the temporal modification of the abundance ratio Mg/Ne in four active regions. They compared the value with the solar value $\mathrm{Mg/Ne}_{\odot}$. The ratio of the two values, β, was then assumed to be indicative of the FIP bias and is plotted versus the time interval since emergence of the active region in Fig. 16. Such strikingly time-proportional modification of abundances was seen for other active regions as well in that study, but had probably gone unnoticed in previous studies because those tended to focus on the brightest and hence youngest active regions. The parts of the active regions that were used for this study tended to be located at the footpoints of the loops, indicating that this is where the fractionation occurs. The authors find that as the loops emerge from beneath the photosphere, they have photospheric composition with the ratio of low-FIP ions to protons being $\sim 1 \times 10^{-4}$, but from then on the material becomes

[1] Vice versa, high FIP elements might be considered depleted with respect to low-FIP elements. See e. g. Wimmer-Schweingruber, 2003 [59] for a brief discussion of this point.

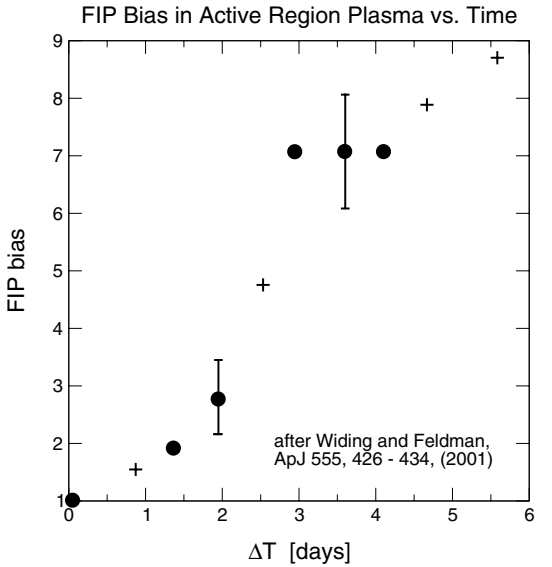


Fig. 16. Enrichment of Mg (as a proxy for low FIP elements) with respect to Ne (as a proxy for high FIP elements) against the time interval since emergence of the active region. After Widing and Feldman (2001) [52].

more and more fractionated at a rate of enrichment of low-FIP ions relative to protons of $\sim 2 \times 10^{-4}$ per day. When loops reconnect with open field lines, the material can escape, feeding into the corona and ultimately the solar wind. When the magnetic configuration is suitable for flux-rope formation and subsequent ejection of coronal[2] mass, such material should be observed in-situ by mass spectrometers. This will be discussed in the following section.

4 Interplanetary Disturbances

4.1 CME Evolution into the Interplanetary Medium

As the CME moves through the corona and out into interplanetary space some of its properties will be modified by the interaction with the corona or the interplanetary medium, or by internal dynamic changes. The flux rope formed during the reconnection events sketched on page 84 will expand and undergo some changes that will determine its properties in interplanetary space. On the one hand, expansion will lead to a cooling of the plasma contained in the CME. The strong magnetic field will lead to an overexpansion of

[2] Some authors prefer to use the term solar here. It is not clear, as underlined by the measurements of Widing and Feldman (2001) [52], that the ejected mass is actually coronal.

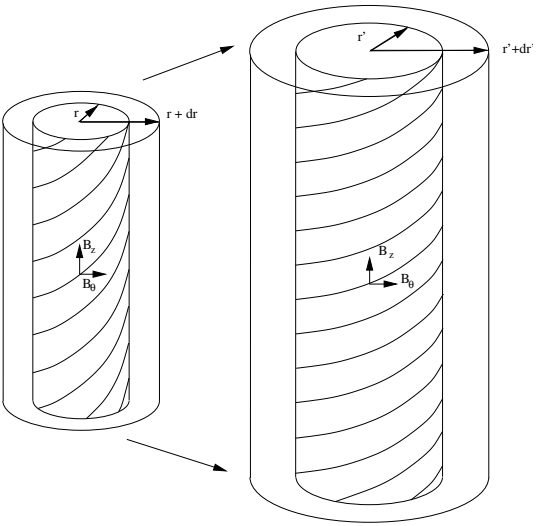


Fig. 17. The expansion of a flux rope conserves the axial and azimuthal fluxes through the annuli between r and dr and r' and dr'. This leads to a more tightly wound flux rope.

the CME relative to the ambient solar wind. This has been observed in interplanetary space, especially at high latitudes (Gosling et al., 1998 [17]), possibly because the solar wind was especially simple during that time period and at those high latitudes. The expansion will also lead to an additional twist of the constituting magnetic field. Because both the axial and azimuthal flux of the magnetic field through an annulus between r and $r+\mathrm{d}r$ must be conserved during expansion (Parker, 1979 [35]), the azimuthal component will increase until the flux rope appears to consist nearly completely of azimuthally wound field. Figure 17 illustrates the process. Of course, this strong winding makes it easier to discern such situations with measurements of the interplanetary magnetic field. Bothmer and Rust (1997) [6] have shown how interplanetary measurements relate to the coronal orientation of the left-behind arcades. Because these will tend to be governed by Hale's law, the orientation of the field will also tend to exhibit a solar-cycle and hemisphere dependent orientation. During even cycles, one expects to first detect a southward component of the field, followed by an eastward and finally a northward turn. Often, the rotation observed in the field appears to contain the heliospheric current sheet (Crooker et al., 1998 [8]).

The topology in interplanetary space in which a large-scale rotation of the magnetic field is observed is called a magnetic cloud. Intriguingly, not all interplanetary counterparts of CMEs are of the magnetic cloud type. Is this simply due to a selection effect, when the spacecraft cuts through the

ICME[3]? Or are CMEs not all of the flux-rope type that models seem to produce, possibly because that's the way we understand these eruptions?

4.2 Interaction with the Interplanetary Medium: The Formation of Shocks

As we have already seen, CMEs leave the Sun with a wide range of speeds which implies that they need to interact with the ambient solar wind. Slow CMEs will tend to slow down the solar wind and be accelerated, fast CMEs will plow into the solar wind and drive a pressure wave infront of them which can steepen into an interplanetary shock when conditions are favorable. CMEs which leave the Sun just at the solar wind speed will be force-free, there are no forces acting on it in the rest frame of the ambient solar wind. Needless to say, this situation is not very frequent.

The faster CMEs will drive compressive disturbances into the interplanetary medium which propagate through it at the fast (or slow) mode speed. Consider a train of two such disturbances, as depicted in Fig. 18. The first compresses and heats the plasma through which it is traveling, leaving behind a hotter plasma. The following disturbance will travel through this plasma

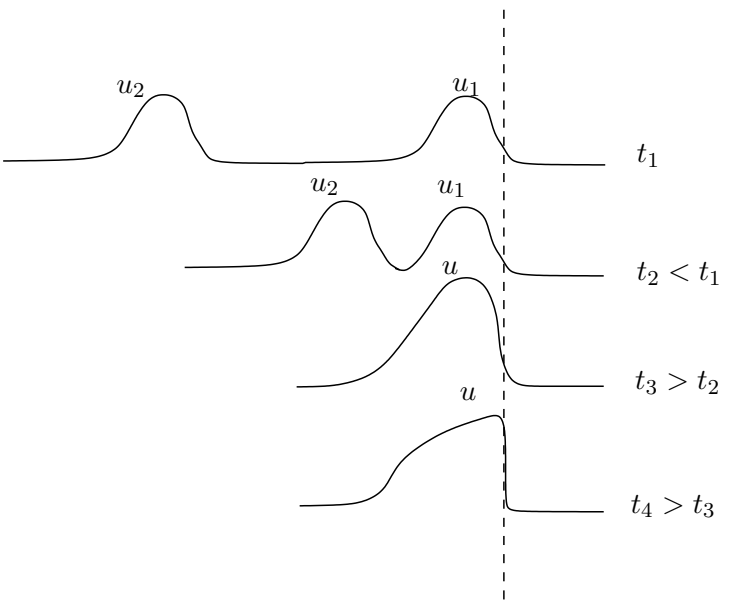


Fig. 18. Shock formation. The trailing disturbance travels at a higher speed than the leading disturbance and catches up. The superposition of many disturbances steepens and turns non linear.

[3] Interplanetary CME, a somewhat strange construction.

at the local fast (or slow) mode speed which is higher than the corresponding speed of the first disturbance because the plasma is hotter and the sound speed has increased. Hence the second disturbance will catch up with the first but cannot overtake it, because it would then run into cooler plasma and slow down. The disturbances all pile up at some leading edge. Gradually this pile-up of disturbances grows, the superposition of a myriad of small disturbances turns into a large disturbance which cannot be described by linearized MHD anymore. The local speed profile gets steeper and steeper until some process dissipates the energy being put into this region. The steep speed profile is thus stabilized by dissipation. This can be achieved e. g. by heat conduction, finite electrical conductivity, viscosity, or dispersion. The first three possibilities are dissipative. They lead to different shock thicknesses. We can imagine a disturbance running through a plasma and steepening to a shock of a thickness determined by the magnetic Reynolds number. The finite electrical conductivity will then decouple magnetic and hydrodynamical fluctuations, and the wave can now only propagate at the speed of sound. If the downstream speed is faster than the speed of sound, the shock is now stabilized. A second disturbance travels at the speed of sound and can't reach the first anymore. If this is not the case, it catches up and another mechanism is needed to stabilize the shock. In this case viscosity will stabilize the shock on a shorter scale, it can be shown that viscosity can stabilize any shock. Heat conduction can only stabilize very weak shocks.

Dispersion can also stabilize shocks. Here the smallest scale length is determined by charge separation and is given by the Debye length. The next largest scales are electron and ion inertial lengths. Again, disturbances will steepen until they reach the longest scale that will stabilize the shock. Since oscillations are the dispersive agent in this case, we expect that energy is dispersed away from the shock in the form of waves. This is indeed the case, dispersive shocks are accompanied by precursor whistler-mode waves.

In order to determine when a shock will be stabilized by dissipation or dispersion requires a careful analysis of the associated scale lengths, the longest scale length that will suffice to stabilize the shock determines the nature of the shock.

We will now go on to describe shocks in the interplanetary medium. The aim is not a complete understanding, but to explain the myriad of names abounding in interplanetary "shockology". Figure 19 defines the geometry we will use in the following few pages. The plasma is approaching the shock from the left-hand side (upstream), at the shock the plasma velocity changes to its downstream value. The angle between the normal to the shock surface (the shock normal) and the magnetic field is generally called θ_{Bn}. There are several quantities that need to be conserved across the shock. For one, an equal amount of matter must flow in from the left as leaves to the right

$$[\rho \boldsymbol{v} \cdot \boldsymbol{n}] = \rho_u \boldsymbol{v}_u \cdot \boldsymbol{n} - \rho_d \boldsymbol{v}_d \cdot \boldsymbol{n} = 0. \tag{32}$$

The square brackets, $[x]$, indicate that the difference between the upstream and downstream quantities is to be taken. We can now estimate the speed at which the shock is moving through interplanetary space,

$$v_s = \frac{\rho_d \boldsymbol{v_d} - \rho_u \boldsymbol{v}_u}{\rho_d - \rho_u} \cdot \boldsymbol{n}. \tag{33}$$

It isn't shock speed per se which determines shock formation, but the component parallel to the magnetic field, $v_{s\|} = v_s / \cos(\theta_{\rm Bn})$. In order for a shock to form, $v_{s\|}$ must exceed v_A, $v_{s\|} > v_A$, in order to overtake the Alfvén waves propagating along the field lines. It isn't necessary for v_s to exceed v_A as long as $\theta_{\rm Bn}$ is large enough. The Alfvén Mach number M_A of the shock is defined as

$$M_A = \frac{v_s - v_u}{v_A}, \tag{34}$$

and gives the shock speed in the upstream frame of reference relative to the Alfvén speed. However, this definition allows fast shocks (propagating at the fast mode speed) to have a Mach number smaller than unity. This counterintuitive situation is circumvented by introducing the critical Mach number

$$M_c = \frac{v_s - v_u}{v_A \cos\theta_{\rm Bn}}. \tag{35}$$

In this notation fast shocks always have $M_c > 1$, the inexistent intermediate ones would have $M_c = 1$, and slow shocks have $M_c < 1$. There are no intermediate shocks because the intermediate mode is not compressive and hence disturbances can't run into each other and steepen to a shock. The definition of the critical Mach number requires knowledge of $\theta_{\rm Bn}$ which can be measured.

A second conserved quantity across the shock is momentum

$$\left[\rho \boldsymbol{v}(\boldsymbol{v}\cdot\boldsymbol{n}) + \left(p + \frac{1}{2}\frac{\boldsymbol{B}^2}{\mu_0}\right)\boldsymbol{n} - \frac{(\boldsymbol{B}\cdot\boldsymbol{n})\boldsymbol{B}}{\mu_0}\right] = 0. \tag{36}$$

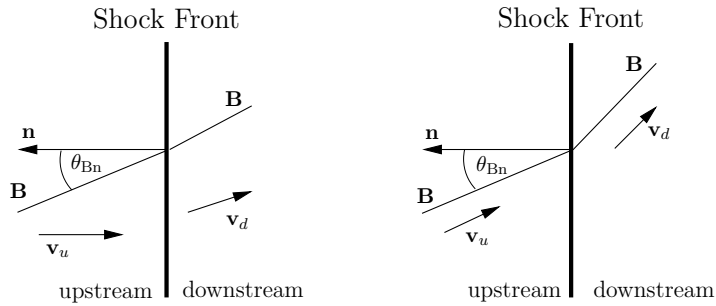


Fig. 19. The geometry of shocks in two different frames of reference: normal incidence (left) and de Hoffmann-Teller (right).

This is a vector equation, the components perpendicular to the shock and parallel to it both satisfy the above relation separately. This implies that one can find a frame of reference, in which the upstream velocity vector $\boldsymbol{v}_u$ is parallel to the shock normal. This is the geometry sketched in the right-hand panel of Fig. 19, the so-called de Hoffmann-Teller frame. Downstream, the velocity points along the magnetic field as well, which means that the induced electric field at the shock surface $\boldsymbol{u} \times \boldsymbol{B}$ vanishes in this frame. The reference frame moves at a speed $\boldsymbol{v}_{\mathrm{dHT}} \times \boldsymbol{B} = -\boldsymbol{E}$.

The next conserved quantity is energy,

$$\left[\boldsymbol{v} \cdot \boldsymbol{n} \left(\frac{\rho \boldsymbol{v}}{2} + \frac{\gamma}{\gamma - 1} p + \frac{\boldsymbol{B}^2}{\mu_0} \right) - \frac{(\boldsymbol{B} \cdot \boldsymbol{n})\,(\boldsymbol{B} \cdot \boldsymbol{v})}{\mu_0} \right] = 0, \tag{37}$$

where $\gamma = c_p/c_v$ is the ratio of specific heats. Moreover, we also have

$$[\boldsymbol{B} \cdot \boldsymbol{n}] = 0, \text{ and} \tag{38}$$

$$[\boldsymbol{n} \times (\boldsymbol{v} \times \boldsymbol{B})] = 0. \tag{39}$$

These conservation equations allow us to derive the downstream quantities based on our knowledge of the upstream quantities of an MHD shock. This is simplified if one first derives the so-called Rankine-Hugoniot relations, which directly relate these quantities

$$\frac{\boldsymbol{v}_u}{\boldsymbol{v}_d} = \frac{1}{X}; \quad \frac{\boldsymbol{B}_u}{\boldsymbol{B}_d} = X \quad \text{etc.}, \tag{40}$$

where, in the case of perpendicular shocks (explained in the following paragraph), X is the solution of

$$2(2 - X)X^2 + [2\beta_u + (\gamma - 1)\beta_u M_u^2 + 2]\gamma X - \gamma(\gamma + 1)\beta_u M_u^2 = 0, \tag{41}$$

where β is the plasma β and M_u the upstream Mach number. For the general case, the equation is more complicated and must in general be obtained numerically. The critical ordering parameter is the upstream Mach number M_u. Note that only upstream quantities appear in (40).

We've already encountered a number of expressions that appear in "shockology", here we give brief explanations of the most common expressions abounding in this field:

fast: A fast shock is formed by fast mode waves steepening into a shock. The fast mode wave is compressive and the magnetic field is deflected away from the shock normal in the downstream region.

slow: A slow shock is formed by a slow mode wave steepening into a shock. The slow mode is compressive and the magnetic field vector is deflected towards the shock normal downstream of the shock.

forward: A forward shock propagates forwards relative to the solar wind.

reverse: A reverse shock propagates backwards in the solar wind frame of reference.

critical: The boundary when resistivity cannot stabilize the shock anymore. Subcritical dissipative shocks are stabilized by resistivity (or the finite electric conductivity), supercritical dissipative shocks by viscosity. The critical fast Mach number is determined by equating the normal component of the downstream shock frame flow speed to the sound speed. Subcritical shocks have slower flow speeds, supercritical ones faster flow speeds.

parallel: A parallel shock has a shock normal parallel to the magnetic field ($\theta_{\mathrm{Bn}} \sim 0$).

perpendicular: a perpendicular shock has a shock normal perpendicular to the magnetic field ($\theta_{\mathrm{Bn}} \sim 90$).

quasi-: all the cases in between are called quasi-parallel ($0 \geq \theta_{\mathrm{Bn}} < 45°$) or quasi-perpendicular ($45 \geq \theta_{\mathrm{Bn}} < 90°$).

dissipative: The shock is stabilized by a dissipative mechanism.

dispersive: The shock is stabilized by a dispersive mechanism.

4.3 Signatures

After this excursion into shock formation, we will briefly discus how ICMEs can be detected in-situ. It is important to notice that the detection of a shock alone does not indicate the presence of an ICME. Shocks may be driven by ICMEs or not (depending on their speed) and the shock may reach an observer while the ICME misses this point in space entirley. ICMEs are not the same beast as interplanetary shocks! Obviously, measurements of the magnetic field topology are an important key to this, however, there are various ways of measuring magnetic topology, and directly measuring the magnetic field is only one of them, as we will see. Table 1 gives a summary of some of the signatures used to detect ICMEs. Strong enhancements in the magnetic field are often an indicator that there is some special magnetic structure passing the spacecraft. If it is accompanied by a long-lasting large spatial rotation of the field, as we expect to pass through when intersecting a flux rope, a magnetic cloud has been detected. Because of the large expansion and related cooling, there is often a very low level of fluctuations in the cloud. Another way of measuring the magnetic topology is to use suprathermal electrons. These electrons stream along the magnetic field and will do so from where they were accelerated. Often CMEs and magnetic clouds are associated with bidirectional electron streaming, a situation which may be due to the connection of the flux rope back to the Sun with its two "feet". The electrons are mirrored at the magnetic constrictions back at the Sun and bounce back and forth along the interplanetary field. Note, however, that bidirectional electron streaming can also mean that the field line is connected back to the Sun on one side and to an interplanetary shock somewhere beyond the spacecraft location. Careful scrutiny of the pitch-angle distribution of the

Table 1. Some signatures of interplanetary CMEs (ICMEs).

Signatures of ICMEs	
Signature	physical. signif.
long B mag. enhancements	strong sol. Field
large spatial temp.	
low level of fluctuations	source?
long-lasting counterstr. e-	connectivity
low kinetic temp. rel. to	
ambient solar wind	source, expansion
a/p enhancement $> 8\%$	source?
charge-state comp.	source, expansion

electrons can help reject such misidentifications. The strong expansion of the ejecta on its way from the Sun to the observer results in a cool plasma in ICMEs. This may also partly be due to the relics of prominence material in the plasma, however, such material is only very rarely seen for reasons that are not understood. Is the filament "emptied out" when the CME erupts? Then what is the bright center of CMEs? Remember that white light tracks the electron density, not neutral particles. Does it, the filament, remain very small and concentrated? In recent years the composition of the solar wind has been established as a new tool for the investigation of the interplanetary plasma. Strong enhancements of the ratio of α particles relative to protons nearly always seem to indicate the presence of an ICME, as does a strange charge-state composition of the heavy solar wind ions. Elevated charge states of iron imply a hot coronal origin, sometimes a puzzling mixture of high and low charge states is observed (Schwenn et al., 1980 [41]; Gloeckler et al., 1999 [14]). How can such a plasma be formed? Equilibrium processes do not result in such mixtures, rapid cooling of strongly heated plasma under extreme non-equilibrium conditions appear to be a likeley candidate (Neukomm and Bochsler, 1996 [31]). Magnetic clouds tend to exhibit higher charge states of O (Henke et al., 1998 [18]) than non-cloud ICMEs. Interestingly, the trailing plasma of ICMEs tends to exhibit the unusual composition or charge states (Wimmer-Schweingruber et al., 1999 [61]). Is this a signature of material from a reconnection area?

4.4 Frequency of CMEs

Coronal mass ejections are the most spectacular consequence of solar activity. These disturbances propagate outwards from the Sun and often drive a strong coronal and interplanetary shock. They are the drivers of gradual solar energetic particle events and accelerate particles throughout interplanetary space. Therefore, it is important to know what their relation to solar activity is. A key study in this respect is the one of Webb and Howard (1994) [51] who investigated the solar cycle variation of coronal mass ejections and the

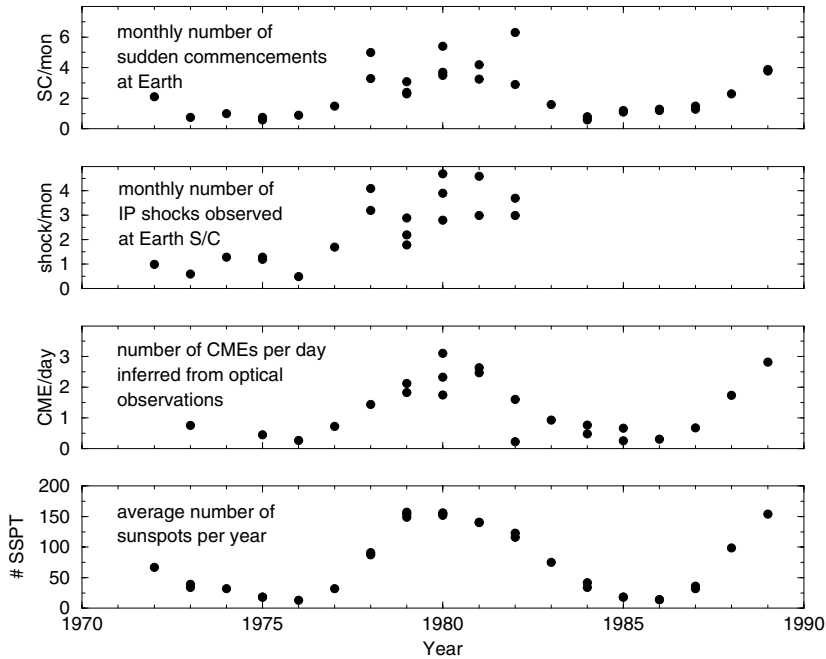


Fig. 20. Time series of yearly averages: monthly number of sudden commencements, monthly number of interplanetary shocks observed at near Earth spacecraft, daily number of CMEs observed by various remote-sensing instruments, yearly sunspot number. Data are from Webb and Howard (1994) [51].

solar wind mass flux. They found that the frequency of coronal mass ejections (CMEs) tends to track the solar activity cycle both in phase and in amplitude, the ratio of CME to solar wind flow tends to track the solar cycle, and that around solar activity maximum, CMEs can provide a significant fraction ($\approx 15\%$) of the average flux to the near-ecliptic solar wind. We have plotted their results in Fig. 20. From top to bottom we have plotted against time (year) the monthly number of sudden commencements at Earth, the monthly number of interplanetary (IP) shocks observed at Earth, the daily occurrence rate of CMEs at the Sun and the yearly averaged sunspot number. The presence of more than one symbol for certain time periods is due to the inclusion of more than one data set. This has the advantage of giving an estimate of the uncertainty of the observed numbers. Obviously, the quantities plotted in the four panels are correlated. We have plotted the daily CME occurrence rate at the Sun vs. yearly averaged sunspot number in Fig. 21. Adding newer data from the LASCO instrument on SOHO (St. Cyr et al., 2000 [49]) does not alter the overall picture but shows some puzzling features. We omit these newer observations. Sudden commencements observed at the Earth are generally accepted to be accompanied by interplanetary shocks (Smith, 1983 [45]).

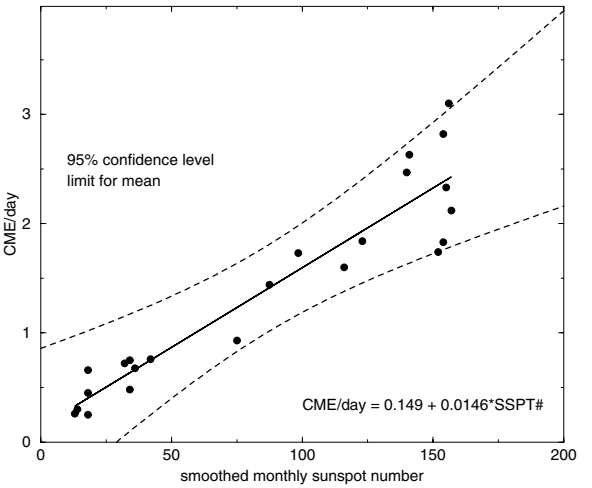


Fig. 21. Daily frequency of coronal mass ejections at the Sun versus smoothed monthly sun spot number. Data are from Webb and Howard (1994) [51]. The solid line shows the best unweighted least-squares linear fit, the dashed curves give the 95% confidence level limits on the expectation value for the number of CMEs per day for a given monthly average sun spot number.

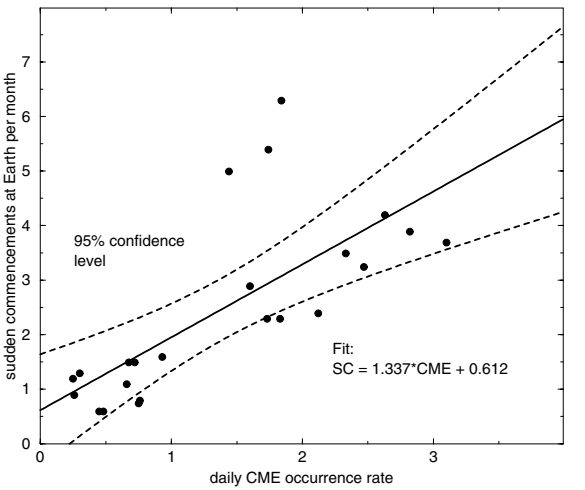


Fig. 22. Monthly number of sudden commencements at Earth versus daily occurrence rate of coronal mass ejections. Data are from Webb and Howard (1994) [51]. The solid line shows the best unweighted least squares linear fit, the dashed curves give the 95% confidence level limits on the expectation value for the number of sudden commencements for a given daily occurrence rate of coronal mass ejections as observed at the Sun.

We have plotted the monthly number of sudden commencements observed at Earth vs. yearly averaged sunspot number in Fig. 22. Other important studies are the one by Lindsay et al. (1994) [26] who investigated the sources of interplanetary shocks at 0.72 AU. This study can serve to give a lower limit on IP shocks at 1 AU. Shea and Smart (1990) [42] compiled a comprehensive summary of all major solar proton events for solar cycles 19, 20, and 21. From their data, we obtain an average occurrence rate of major proton events of $0.085 \times$ sunspot number per year.

From these investigations we learn that the number of coronal mass ejections impinging on the Earth increases with increasing solar activity. The number of sudden commencements increases linearly with the observed solar CME rate. Thus we may conclude that for elevated solar activity the influence of coronal mass ejections will increase. For the case of lower solar activity in the past the converse is true, we would expect less CMEs to disturb the close environment of the Earth.

4.5 Global Merged Interaction Regions

The latitudinal and longitudinal extent of CMEs is large, one typical CME can span a substantial fraction of the surface of a sphere surrounding the Sun. Their source longitudes are uniformly spread about the Sun, their source latitudes are strongly solar activity cycle dependent. During activity maximum, CMEs can be ejected from a large latitudinal extent, this largely reflects the wide latitudinal excursions of the current sheet or streamer belt. As we have seen in the previous talk [38], active regions tend to be distributed nearly over the entire Sun during activity maximum. During activity minimum the rarer ejections tend to concentrate along the streamer belt. Let us consider a much simplified activity maximum situation where CMEs are ejected with equal probability over the entire surface of the Sun. How probable is it that a CME will be ejected from a source region which lies within the cone opening angle of a previous CME? Let α be the half angle of the cone, R be the CME rate at the Sun in CMEs per day. The solid angle spanned by the cone is just $(1 - \cos\alpha)/2$. We start off with a CME being ejected at some location on the Sun. Then the probability of having at least one CME erupt from a source region within the cone of the first CME in the course of d days is given by the simple binomial

$$\begin{aligned} P &= \sum_{i=1}^{d\,R} \binom{d\,R}{i} \left(\frac{1-\cos\alpha}{2}\right)^{i} \left(\frac{1+\cos\alpha}{2}\right)^{d\,R-i} \\ &= 1 - \binom{d\,R}{0} \left(\frac{1+\cos\alpha}{2}\right)^{d\,R} = 1 - \left(\frac{1+\cos\alpha}{2}\right)^{d\,R} . \end{aligned} \tag{42}$$

This probability grows very fast, as can be seen from Fig. 23. During activity maximum periods, there is a 50% chance of a CME to be ejected into the cone

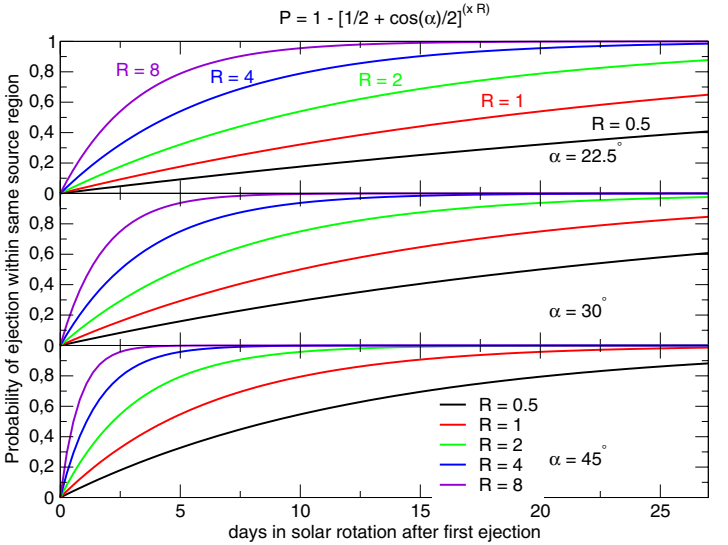


Fig. 23. Probability of ejecting a CME within the conic extent of an initial CME ejected on day 0. The top panel shows the case for cone half width $\alpha = 22.5°$, the middle panel for $\alpha = 30°$ and the bottom panel for $\alpha = 45°$. The different curves in each panel reflect different ejection rates ranging from 0.5 CMEs per day at solar activity minimum to 8 per day at solar activity maximum.

opening of the first CME within about one day, during activity minimum, this time period is much longer, more than half a solar rotation. Given the wide speed distribution of CMEs and of the ambient solar wind with which they interact there is a high probability for a fast CME to catch up with a slow one and to interact with it within the heliosphere at some heliocentric distance s. Equating travel times one trivially finds the expected result that s is inversely proportional to the speed difference $v_f - v_s$ and proportional to the time difference Δt between ejection,

$$s = \frac{v_s\, v_f}{v_f - v_s}\, \Delta T.$$

Given a speed distribution of CMEs one can now easily compute the probability that a CME will interact with another inside a heliocentric distance $s\ldots$

The interaction of ICMEs will lead to giant regions of enhanced magnetic field and with complicated magnetic topology. During activity maximum it is quite probable that a substantial fraction of the full sphere will be covered by CMEs ejected in the course of a few solar rotations. Thus CMEs will tend to form giant shells in the outer heliosphere, called global merged interaction regions, GMIRs. These GMIRs serve as effective barriers against galactic

cosmic rays. These can penetrate the heliosphere more easily during solar activity minimum than during maximum conditions.

5 Particle Acceleration and Transport

5.1 Particles in Electromagnetic Fields

The motion of particles in electromagnetic field can be described in various ways, depending on the properties of the ambient plasma. In plasmas in which the magnetic field varies smoothly and variations are small, or conversely, the scale length of the system is much larger than the Larmor radius of the particles, often adiabatic invariants, guiding center approximations, and drift properties are used. Such a description is often used to explain cosmic ray drifts in the heliosphere. If there are large-scale variations, particle trajectories are calculated by integrating the equations of motion in the background plasma. This method is often used to describe particles in quasi-perpendicular shocks. Finally, when the plasma is turbulent and strong scattering of the particles is expected, transport equations including pitch-angle scattering and other processes are used.

While the problem of describing a particles motion in a magnetic field sounds rather simple, there is more to it than meets the eye. Many interesting phenomena can be understood by investigating the equation of motion of a particle of charge q (expressed in elementary charges e) and rest mass m in a magnetic field $\boldsymbol{B}$.

$$\frac{\mathrm{d}\boldsymbol{p}}{\mathrm{d}t} = q\,(\boldsymbol{E} + \boldsymbol{v} \times \boldsymbol{B})\,, \tag{43}$$

where $\boldsymbol{p} = \gamma\, m\, \boldsymbol{v}$ is the relativistic momentum and

$$\gamma = \frac{1}{\sqrt{1 - v^2 \,/\, c^2}}. \tag{44}$$

The energy of a particle is given by $U = m\gamma c^2$. If the electric field disappears everywhere, $\boldsymbol{E} = 0$, then the induction law implies that $\boldsymbol{B}$ is time independent, $\dot{\boldsymbol{B}} = 0$. In this case we can take the scalar product of the equation of motion with momentum $\boldsymbol{p}$,

$$\frac{\mathrm{d}\boldsymbol{p}}{\mathrm{d}t} \cdot \boldsymbol{p} = q\,(\boldsymbol{v} \times \boldsymbol{B}) \cdot \boldsymbol{p} = 0, \tag{45}$$

because velocity $\boldsymbol{v}$ is parallel to momentum $\boldsymbol{p}$. This implies that the magnitude of momentum is conserved in any magnetic field, $\mathrm{d}|\boldsymbol{p}| \,/\, \mathrm{d}t = 0$, as is total kinetic energy U and the Lorentz factor γ. A magnetic field does not perform any work. We can write momentum as $\gamma m \boldsymbol{v}$ in the equation of motion to determine the acceleration of the particle

$$\frac{\mathrm{d}\boldsymbol{v}}{\mathrm{d}t} = \frac{q}{m\,\gamma}\,(\boldsymbol{v} \times \boldsymbol{B})\,. \tag{46}$$

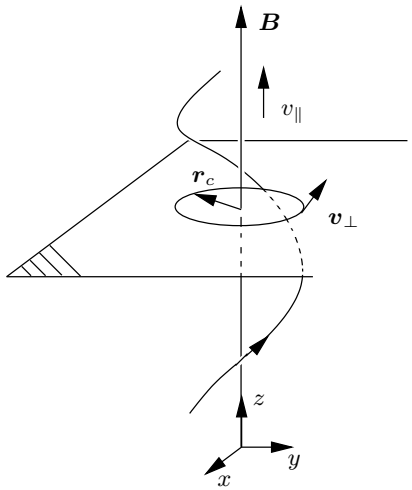


Fig. 24. The motion of a charged particle in a magnetic field can be viewed as a superposition of a circular motions perpendicular to the magnetic field and a translation along, parallel to, the field.

Next, we define a new quantity $\boldsymbol{\Omega}$

$$\boldsymbol{\Omega} = \frac{-q}{m\,\gamma}\boldsymbol{B} \tag{47}$$

which allows us to rewrite the previous equation (46)

$$\frac{\mathrm{d}\boldsymbol{v}}{\mathrm{d}t} = \boldsymbol{\Omega} \times \boldsymbol{v}. \tag{48}$$

$\boldsymbol{\Omega}$ is the so-called gyro-frequency, in the non-relativistic approximation $v^2\,/\,c^2 \ll 1$ this quantity tends towards the cyclotron frequency.

Next, we divide the particles motion into a motion parallel and perpendicular to the magnetic field, as illustrated in Fig. 24. We define the z component of our frame of reference parallel to $\boldsymbol{B}$ and can now rewrite (48) in components

$$\begin{aligned} \dot{v}_x &= \Omega_y\,v_z - \Omega_z\,v_y = -\Omega_z\,v_y, \\ \dot{v}_y &= \Omega_z\,v_x - \Omega_x\,v_z = \Omega_x\,v_z, \\ \dot{v}_z &= \Omega_x\,v_y - \Omega_y\,v_x = 0. \end{aligned} \tag{49}$$

because $\boldsymbol{\Omega}$ points along $\boldsymbol{B}$, $\Omega_x = \Omega_y = 0$, which implies that $\dot{v}_z = 0$, i. e. a regular motion along z. Hence we can describe the motion of a particle as a superposition of a circular motion perpendicular to the magnetic field and a linear motion along it.

$$\boldsymbol{v}_\perp = \boldsymbol{\Omega} \times \boldsymbol{r}_c, \tag{50}$$

where $\boldsymbol{r}_c$ is the position vector of the particle as seen from the field line which it is circling, or, in other words, as seen from an imaginary point which lies in the middle of the gyration motion and moves along the magnetic field line with a speed $v_\parallel$. This point is called "guiding center", r_c is called the gyro-radius. In the non-relativistic case this is called the Larmor radius,

$$\boldsymbol{r}_c = \frac{\boldsymbol{v} \times \boldsymbol{B}}{\Omega^2} = \frac{m\,\gamma}{q} \frac{\boldsymbol{B} \times \boldsymbol{v}}{B^2} = \frac{1}{q} \frac{\boldsymbol{B} \times \boldsymbol{p}}{B^2}. \tag{51}$$

The concept of a "guiding center" allows one one to generalize the previous discussion to the case of a non-homogeneous magnetic field by dividing the motion into a motion parallel and perpendicular to the magnetic field. The guiding center motion is the parallel motion superimposed with a drift motion $\boldsymbol{v}_D$ perpendicular to the field. In addition, there is the perpendicular motion due to the gyration of the particle, $\boldsymbol{Omega} \times \boldsymbol{r}_c$,

$$\boldsymbol{v} = \boldsymbol{v}_\parallel + \boldsymbol{v}_\perp = \boldsymbol{v}_\parallel + \boldsymbol{v}_D + \omega r = \boldsymbol{v}_{\mathrm{gc}} + \boldsymbol{Omega} \times \boldsymbol{r}_c. \tag{52}$$

Averaging over a gyroperiod removes the term $\boldsymbol{Omega} \times \boldsymbol{r}_c$ and the motion of the article is described by the motion of the guiding center. The particle always moves within a gyroradius of the guiding center. We can evaluate r_c for the field configuration at hand,

$$r_c = |\boldsymbol{r}_c| = \frac{1}{q\,B^2} \left(B_z^2\, p_y^2 + B_z^2\, p_x^2\right)^{1/2} = \frac{|p_\perp|}{q\,B} = \frac{p \sin\alpha}{q\,B}, \tag{53}$$

where the angle α is defined by $\tan\alpha = p_\perp / p_\parallel$ and is called pitch angle. For a circular motion, (i. e. $\alpha = \pi/2$) we have $r_c = p/(qB)$. The quantity

$$c\,B\,r_c = \frac{p\,c}{q} \tag{54}$$

is called magnetic rigidity and has units Volt, as is easily verified. It describes the resistance to a change of direction, the more rigid a particle, the more it resists a force that wants to change its trajectory. Sometimes, only the momentum component perpendicular to the magnetic field is used for this definition. Then the Larmor radius can be written as $r_L = P/B$.

Next, we investigate the situation where the magnetic field is not constant, but varies slowly on a scale length L

$$\frac{1}{L} \sim \left| \frac{1}{B} \frac{\partial\, B_i}{\partial x_j} \right|, \tag{55}$$

i. e. $1/L$ corresponds to the maximum of $|(1/B)(\partial B/\partial x_j)|$. In the following, we assume that L is always much larger than the distance the particle traverses during one gyration period, $\tau = 2\pi/\Omega$,

$$L \gg v\tau \gg r_c. \tag{56}$$

This implies that the magnetic field does not change appreciably inside a gyroradius and that we may use its value at the guiding center instead of the instantaneous value at the particle itself for our calculations. In this approximation, all results will be accurate to order $\pm(r_c/L)^2$ or $\pm(v\tau/L)^2$. Once the position of the particle is given, the position of the guiding center is

$$\boldsymbol{x}_G = \boldsymbol{x} - \boldsymbol{r}_c, \tag{57}$$

and its velocity is

$$\begin{aligned} \boldsymbol{V}_G &= \frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}t} - \frac{\mathrm{d}\boldsymbol{r}_c}{\mathrm{d}t} = \boldsymbol{v} - \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\boldsymbol{B}\times\boldsymbol{p}}{q\,B^2}\right), \\ &= \boldsymbol{v} - \frac{1}{q}\left[\frac{\mathrm{d}\boldsymbol{p}}{\mathrm{d}t}\times\frac{\boldsymbol{B}}{B^2} + \boldsymbol{p}\times\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\boldsymbol{B}}{B^2}\right)\right], \end{aligned} \tag{58}$$

where we have inserted the expression for the gyroradius, (51).

Drifts of particles in electromagnetic fields results from changes in their gyroradius in the course of a gyration. This can happen when their speed changes (e. g. in a $\boldsymbol{E}\times\boldsymbol{B}$ drift) or when the magnetic field changes (e. g. in a ∇B drift). Particles with pitch angle 0° do not experience drifts, except for curvature drift, which stems from a parallel motion.

Averageing over a gyroperiod, (58) can be rewritten[4]

$$\boldsymbol{V}_{\mathrm{G}\perp} = \frac{pv}{qB}\left[\frac{1}{2}\sin^2\alpha\frac{\boldsymbol{B}\times\boldsymbol{\nabla}B}{B^2} + \cos^2\alpha\frac{\boldsymbol{B}\times[(\boldsymbol{B}\cdot\boldsymbol{\nabla})\,\boldsymbol{B}]}{B^3}\right]. \tag{59}$$

We need to note that this equation is not correct in all conceivable field configurations. It is only valid for the assumption that we are allowed to replace the field $\boldsymbol{B}$ at the particles location by the field $\boldsymbol{B}$ at the guiding center. This expression is correct to second order in r_c/L and $v\tau/L$ and that is all. It is **not** a generally valid equation such as the equation of motion, (43).

However, even if the previous expression is not valid in all cases, it does tell us alot about the motion of the guiding center in a smooth magnetic field. The first expression in the square brackets describes the influence of a transversal gradient in the field. The second term results from the curvature of the field lines. The first term can be rewritten in a more conventional form by inserting the expression for the gyroradius and by considering the pitch angle, $v_\perp = v\sin\alpha$.

$$\boldsymbol{V}_{\mathrm{G}\perp\nabla} = \frac{1}{2}r_c v_\perp\frac{\boldsymbol{B}\times\boldsymbol{\nabla}B}{B^2}. \tag{60}$$

This expression describes a drift perpendicular to the direction $\boldsymbol{B}\times\boldsymbol{\nabla}B$, i. e. perpendicular to $\boldsymbol{B}$ and $\boldsymbol{\nabla}B$. It points in opposite directions for particles of opposite charge because charge enters the equation as an odd power. An important example for this drift is the modulation of galactic cosmic ray ions

[4] The derivation is not quite straightforward.

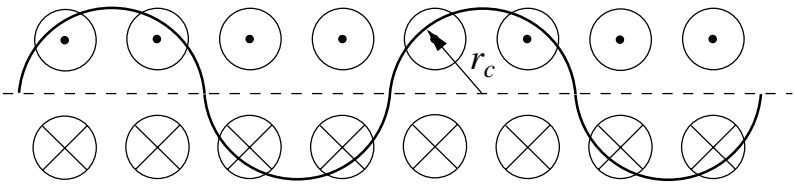


Fig. 25. ∇B drift in a strongly oversimplified configuration of the heliospheric current sheet.

and electrons. Let us consider a dramatic oversimplification of the heliospheric current sheet as sketched in Fig. 25. Some particles drift along the current sheet, positively charged ones drift one way, negatively charged ones the other way. This could explain the differences measured in GCR ions and electrons. However, our assumption that the field does not change on scales large compared to the gyroradius is severely violated in this sketch and in reality.

The second term in square brackets in (59) can be rewritten using $v_\parallel = v\cos\alpha$. Thus we obtain

$$\boldsymbol{V}_{\mathrm{G}\parallel\mathrm{curv}} = \frac{\gamma m v_\parallel^2}{qB^4}\left[\boldsymbol{B}\times\{(\boldsymbol{B}\cdot\boldsymbol{\nabla})\,\boldsymbol{B}\}\right], \tag{61}$$

which again may be more familiar. This curvature drift points perpendicular to a surface area which contains the surface spanned by the curved field and points in opposite directions for oppositely charged particles.

As a particle moves along its gyration radius in a magnetic field, it induces a current $I = q\omega/2\pi$ in the inside of the circle. The particles magnetic moment is defined as the product of the induced current and the area spanned by the circle,

$$\begin{aligned}\boldsymbol{\mu} &= \frac{1}{2}\int d^3x\boldsymbol{x}\times\boldsymbol{I} = \frac{1}{2}\int \boldsymbol{x}\times\boldsymbol{dl}I \\ &= \frac{1}{2}\oint r_L dl\boldsymbol{n}I = \boldsymbol{n}Ir_L^2\pi \qquad (62)\\ &= \boldsymbol{n}\frac{q\omega}{2\pi}\left(\frac{mv_\perp}{qB}\right)^2\pi \\ &= \boldsymbol{n}\frac{q^2B}{2\pi m}\left(\frac{mv_\perp}{qB}\right)^2\pi \\ &= \frac{mv_\perp^2}{2}\frac{\boldsymbol{n}B}{B^2}, \qquad (63)\end{aligned}$$

hence

$$\boldsymbol{\mu} = -\frac{mv_\perp^2}{2}\frac{\boldsymbol{B}}{B^2}. \tag{64}$$

The magnetic moment is independent of the particles charge and points in the direction opposite to the magnetic field. A plasma is diamagnetic.

5.2 Shock Acceleration

So how can a magnetic field accelerate a particle? If the magnetic field is frozen in to the plasma the electric conductivity must be extremely large and hence there can be no electric fields in the plasma. Now we all know that magnetic fields do not perform work on particles, so again, how can particles be accelerated in the interplanetary medium?

There are three different types of particle acceleration in the interplanetary medium,

- scatter-free acceleration in the electric field induced at the shock front, also called shock-drift acceleration,
- stochastic acceleration in turbulent media,
- acceleration by multiple reflections in the plasma parcels converging on the shock front, this is called diffusive acceleration.

Scatter-Free or Shock-Drift Acceleration

The simplest situation arises in quasi-perpendicular shocks, when the electric field $\boldsymbol{E} = -\boldsymbol{u}_u \times \boldsymbol{B}_u$ induced at the shock front is maximal. The electric field points along the shock front and the particles drift along the shock, see Fig. 26. The longer they drift along the shock, the more energy they gain. Their energy gain is proportional to the product of the particles charge, the induced electric field, and the length of the drift path in the shock. As the particle leaves the shock, it feels the magnetic field, and is bent back into the shock region. For low levels of turbulence this motion is more or less scatter free and the particle can cross the shock several times, always gaining energy. In fast magnetosonic and quasi-perpendicular shocks (a large fraction of the shocks outside 1 AU), the particles will also feel drifts in the shock frame. That motion can be divided into two components, the ∇B-drift and curvature drift due to the curvature of the shock. In fast quasi-perpendicular shocks ∇B-drift dominates and the particles gain energy. In slow shocks, the gradient points the other way and it is curvature drift that is parallel to the electric field. As the particles gain energy, their Larmor radii will gradually increase and at some point they will be scattered away from the shock, thus escaping from it. At this point, the particle has gained a substantial amount of energy. Particles with a small velocity component relative to the shock in the shock frame are the ones that remain stuck to the shock longest and hence gain the most energy. Thus, the initial velocity vector and pitch angle are important factors in determining whether particles gain remain in the shock for an appreciable amount of time and gain energy. The spectrum of accelerated particles is not easily derived, but must be determined by detailed

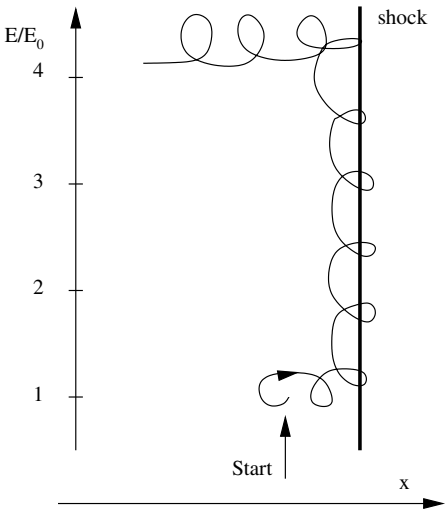


Fig. 26. Shock-drift acceleration. Depending on its initial conditions, the particle is accelerated by the induced electric field along the shock surface.

calculations of many particle trajectory calculations. The trajectories or final energy gains need to be weighted by the distribution in initial conditions. At high energy, a power law in momentum results – we will soon see that this is inherent to the process of shock acceleration.

Not all particles are injected into the acceleration process with the same efficiency. This process readily accepts particles whose velocity vector already points along the shock. This is the case for so-called pick-up ions which are very efficiently accelerated by shock-drift because they form a suprathermal source population.

Scattering in the Interplanetary Medium

Stochastic and diffusive acceleration both rely on scattering in the interplanetary medium. Let us simplify this process by a reflection off a moving wall, as sketched in Fig. 27. If the wall moves anti parallel to the particle at a speed U, the particle gains energy. To understand this, we need to transform into the rest frame of the wall. In this system, the particle will be reflected, a mere change in momentum will occur, and energy will be conserved, $E_1 = E_2$. After this reflection, we transform back into the original frame of reference and we can calculate the energy gain. This is easiest done using the relativistic notation of the four-vector $(E, \boldsymbol{p})$. We denote the quantities in the frame of reference of the wall with primes, then

$$\begin{pmatrix} E_1' \\ p_1' \end{pmatrix} = \begin{pmatrix} \gamma_1 & -\gamma_1\beta_1 \\ -\gamma_1\beta_1 & \gamma_1 \end{pmatrix} \begin{pmatrix} E_1 \\ p_1 \end{pmatrix}, \tag{65}$$

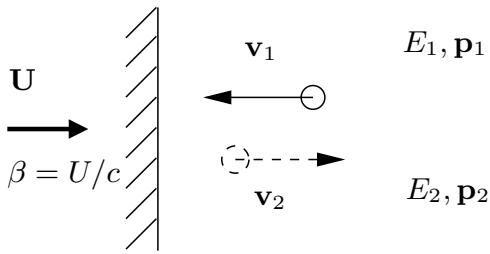


Fig. 27. Energy gain by reflection.

$$\begin{pmatrix} E_2' \\ p_2' \end{pmatrix} = \begin{pmatrix} \gamma_2 & -\gamma_2\beta_2 \\ -\gamma_2\beta_2 & \gamma_2 \end{pmatrix} \begin{pmatrix} E_2 \\ p_2 \end{pmatrix}, \tag{66}$$

where $\beta_1 = U/c$, $\beta_2 = -U/c$,and $\gamma_i = 1/\sqrt{1-\beta_i^2}$. In the frame of reference of the wall $p_2 = -p_1$ and $E_1' = E_2'$ and hence

$$\begin{pmatrix} E_2' \\ p_2' \end{pmatrix} = \begin{pmatrix} \gamma_1 & -\gamma_1\beta_1 \\ \gamma_1\beta_1 & -\gamma_1 \end{pmatrix} \begin{pmatrix} E_1 \\ p_1 \end{pmatrix}. \tag{67}$$

Next we transform back into the original frame of reference

$$\begin{aligned} \begin{pmatrix} E_2 \\ p_2 \end{pmatrix} &= \begin{pmatrix} \gamma_2 & -\gamma_2\beta_2 \\ -\gamma_2\beta_2 & \gamma_2 \end{pmatrix} \begin{pmatrix} \gamma_1 & -\gamma_1\beta_1 \\ \gamma_1\beta_1 & -\gamma_1 \end{pmatrix} \begin{pmatrix} E_1 \\ p_1 \end{pmatrix}, \\ &= \begin{pmatrix} \gamma_1\gamma_2 - \gamma_1\gamma_2\beta_1\beta_2 & -\gamma_1\gamma_2\beta_1 + \gamma_1\gamma_2\beta_2 \\ -\gamma_1\gamma_2\beta_2 + \gamma_1\gamma_2\beta_1 & \gamma_1\gamma_2\beta_1\beta_2 - \gamma_1\gamma_2 \end{pmatrix} \begin{pmatrix} E_1 \\ p_1 \end{pmatrix}. \end{aligned} \tag{68}$$

Therefore, we have

$$E_2 = (\gamma_1\gamma_2 - \gamma_1\gamma_2\beta_1\beta_2)\, E_1 + (-\gamma_1\gamma_2\beta_1 + \gamma_1\gamma_2\beta_2)\, p_1. \tag{69}$$

Because of $\beta_2 = -\beta_1 = \beta$, $\beta_{\text{particle}} = p_1/E_1$, and $\gamma_1 = \gamma_2 = \gamma$

$$\begin{aligned} E_2 &= \left(\gamma^2 + \gamma^2\beta^2\right) E_1 + 2\gamma^2\beta p_1, \\ \frac{E_2}{E_1} &= \frac{1 + 2\beta p_1/E_1 + \beta^2}{1-\beta^2}, \\ \frac{E_2}{E_1} &= \frac{1 + 2\beta\beta_{\text{particle}} + \beta^2}{1-\beta^2}. \end{aligned} \tag{70}$$

Because $\beta^2 > 0$ the particle will gain energy. In an isotropic turbulent medium, the scattering centers are isotropically distributed, and the particle will be more likely to hit a wall head on and gain energy than to catch up with one and loose energy. You pass less cars on the highway than you see cars heading towards you on the other side of the lane (at least if you're driving on the right side of the road...). This acceleration process was first proposed by Fermi (1949) [12] and is therefore called Fermi acceleration.

Stochastic Acceleration

Particles are accelerated in the turbulent medium behind (downstream of) the shock. The waves generated by the shock and the accelerated particles have wave vectors in all directions. Particles gain or loose energy depending on how they encounter the wave. Because it is more likely that they run into a wave (and gain energy) than catch up with it (and loose energy), a net energy gain results from this process called second-order Fermi acceleration. We will now derive the spectrum of stochastically accelerated particles using the equation of continuity in momentum space.

$$\frac{\partial}{\partial p}\left[\dot{p}f(p)\right] + \frac{f(p)}{T} = Q\delta(p - p_0), \tag{71}$$

where $f(p)$ is the distribution function in momentum space, T the average time a particles spends in the acceleration process (or region), $\dot{p}$ the average momentum gain rate, and Q a source term. As a simplification we assume that the particles acquire a momentum increment that is proportional to the momentum they already have, cf. (70), i. e. $\dot{p} = \alpha p$. Inserting into (71) we obtain

$$\frac{\partial}{\partial p}\left[\alpha p f(p)\right] + \frac{f(p)}{T} = Q\delta(p - p_0). \tag{72}$$

A rigid derivation of the problem would require carrying along a second order derivative term which describes diffusion in momentum space, hence the name "stochastic" acceleration. However, the result is not changed dramatically if we neglect this term, so will do so for simplicities sake. Following Jones (1994) [20], we thus treat the simpler case of an ordinary first order differential equation which is easily solved

$$\begin{aligned} f &= e^{-\int dp \frac{\left(1+\frac{1}{\alpha T}\right)}{p}} \int dp \frac{Q}{\alpha p}\delta(p - p_0) e^{\int dp \frac{\left(1+\frac{1}{\alpha T}\right)}{p}}, \\ &= e^{-\left(1+\frac{1}{\alpha T}\right)\ln|p|} \int dp \frac{Q}{\alpha p}\delta(p - p_0) e^{\ln|p|\left(1+\frac{1}{\alpha T}\right)}, \\ &= p^{-\left(1+\frac{1}{\alpha T}\right)} \frac{Q}{\alpha p_0} p_0^{\left(1+\frac{1}{\alpha T}\right)}, \\ &= \frac{Q}{\alpha p_0}\left(\frac{p}{p_0}\right)^{-\left(1+\frac{1}{\alpha T}\right)}. \end{aligned} \tag{73}$$

Thus we obtain a power law in momentum for the distribution function as long as the product αT is independent of momentum p. Of course, we don't know whether this is actually the case. Observations show that $\alpha T \sim 1$ for unknown reasons. Stochastic acceleration can not accelerate particles up to very high energies, such as the several MeV/amu observed in interplanetary space, however, it appears to be an important process for pre-acceleration of particles to several tens to few hundreds of keV/amu. Therefore, stochastic acceleration is often considered an important mechanism to

provide the source population for further acceleration in the more effective diffusive shock acceleration.

Stochastic acceleration does not necessarily need to occur at interplanetary shocks. Observations of the energy spectra of low-energy pick-up ions in corotating interaction regions (CIRs) show that these particles are accelerated in the turbulent regions between the forward-reverse shock pairs of CIRs, not at the shocks themselves (Gloeckler et al., 1994 [15]). In fact, not even shock pairs are required, turbulent regions, defined by regions of high $\delta B/B$, are efficient in accelerating suprathermal particles to higher energies (Schwadron et al., 1996 [40]).

Diffusive Acceleration

More efficient acceleration arises when a particle has enough energy to cross back and forth from one side of a quasi-parallel shock to the other. If it scatters back and forth, it will gain energy, as can be seen from the following considerations. Consider a particle that is energetic enough to pass from downstream of the shock to the upstream region. There it is reflected by scattering and returns downstream of the shock accompanied by a gain in energy because, in the shock frame, the medium is moving towards the shock at the shock speed. The same process occurs in the downstream region. There the particle will be reflected by disturbances moving towards the shock too. These disturbances are the ones that slowly built up to drive the shock in the first place. Because the plasma is hotter behind the shock, they will always catch up with the shock and hence always be faster than the shock. Hence the particle sees converging scattering centers at the shock.This process is called first-order Fermi acceleration. If the particle is scattered at an angle such that its velocity component back towards the shock is sufficient to cross the shock into the upstream region again, the cycle repeats.

In contrast to stochastic acceleration, where the particle is accelerated by the turbulence *on one side of the shock*, particles are accelerated *on both sides of the shock* in diffusive acceleration. They can cross the shock many times and every time they gain a momentum increment $\delta p \propto p$. Obviously, this process is much more efficient than stochastic acceleration. Again following Jones (1994) [20] we follow a particle through its history of shock crossings. After N cycles (consisting of two shock crossings), the particle will have momentum

$$p(N) = p_0 \prod_{i=1}^{N} \left(1 + \left[\frac{\delta p}{p}\right]_i\right). \tag{74}$$

The term in square brackets could be momentum dependent, therefore, we have given it an index i. During each cycle, there is a small, but non-vanishing probability ϵ_i that the particle will not be reflected downstream of the shock and will escape from the acceleration region. Hence, the probability for the

particle to survive N cycles of the process is

$$\mathcal{P}(N) = \prod_{i=1}^{N} (1 - \epsilon_i) \,. \tag{75}$$

We take the logarithm of (74) and (75)

$$\begin{aligned} \ln\left[\frac{p(N)}{p_0}\right] &= \sum_{i=1}^{N} \ln\left(1 + \left[\frac{\delta p}{p}\right]_i\right) \\ &\approx \sum_{i=1}^{N} \left[\frac{\delta p}{p}\right]_i , \end{aligned} \tag{76}$$

$$\begin{aligned} \ln\left[\mathcal{P}(N)\right] &= \sum_{i=1}^{N} \ln\left(1 - \epsilon_i\right) \\ &\approx -\sum_{i=1}^{N} \epsilon_i , \end{aligned} \tag{77}$$

and take the ratio

$$\frac{\ln \mathcal{P}(N)}{\ln\left[p(N)/p_0\right]} = \frac{-\sum_{i=1}^{N} \epsilon_i}{\sum_{i=1}^{N} \left[\frac{\delta p}{p}\right]} = -\Gamma(N?), \tag{78}$$

from which follows

$$\mathcal{P}(N) = \left(\frac{p}{p_0}\right)^{-\Gamma(N?)} . \tag{79}$$

This expression already tells us the probability of the particle to have momentum p. The question mark after N is supposed to remind us that we don't know whether the exponent Γ depends on N or not. If it does, (79) is not a power law in momentum p.

The situation is similar to that of stochastic acceleration. Let us define τ_i as the time needed for a cycle i. Then we can divide the denominator and the numerator of (78) by the total time spent in the acceleration process,

$$\frac{\sum_{i=1}^{N} \left[\frac{\delta p}{p}\right]_i}{\sum_{i=1}^{N} \tau_i} = \frac{1}{p}\frac{\mathrm{d}p}{\mathrm{d}t} = \alpha(N), \tag{80}$$

$$\frac{\sum_{i=1}^{N} \epsilon_i}{\sum_{i=1}^{N} \tau_i} = \frac{\mathrm{P(loss)}}{\mathrm{time}} = \frac{1}{T(N)}. \tag{81}$$

Next we rewrite $\Gamma(N?)$ from (78) using the average relative momentum gain $\alpha(N)$ and loss time $T(N)$

$$\Gamma(N) = \frac{1}{\alpha(N) T(N)} \tag{82}$$

to find the spectrum of accelerated particles

$$f(P) \propto \left(\frac{p}{p_0}\right)^{\frac{-1}{\alpha(N)T(N)}}, \tag{83}$$

as in stochastic acceleration, if the product $\alpha(N)T(N)$ is independent of N or momentum p. In order to find out whether it is, we need to investigate the momentum gain in more detail. One can show (see Jones, 1994 [20] for references), that the rate of momentum change is given by

$$\dot{p} = -\frac{p}{3}\boldsymbol{\nabla}\cdot\boldsymbol{u}, \tag{84}$$

where $\boldsymbol{u}$ is the flow velocity of the plasma. Because only the normal component (to the shock surface, i. e. parallel to the shock normal) is important, we define it as the x-direction and write

$$\dot{p} = -\frac{p}{3}\frac{\partial u}{\partial x}. \tag{85}$$

This can be integrated along a particle trajectory from upstream to downstream

$$\begin{aligned}\delta p = -\frac{1}{3}\int_u^{\mathrm{d}} p\left(\frac{\mathrm{d}u}{\mathrm{d}x}\right)\mathrm{d}t \;&=\; -\frac{1}{3}\int_u^{\mathrm{d}} p\left(\frac{\mathrm{d}u}{\mathrm{d}x}\right)\frac{\mathrm{d}x}{v_x},\\ &= \frac{1}{3}\frac{p}{v_x}(u_u - u_d).\end{aligned} \tag{86}$$

to describe the momentum gained by a particle. The average momentum gain for all particles traversing the shock can be computed by averaging over the flux through the shock. Assuming the particles to be isotropic,

$$\left[\frac{1}{v_x}\right] = \frac{2}{v}, \tag{87}$$

and because a cycle consists of two shock crossings, we have

$$\left[\frac{\delta p}{p}\right]_i = \frac{4}{3}\frac{u_u - u_d}{v_i}. \tag{88}$$

In order to find the particle spectrum, we now need to find the probability that the particle will actually return to the shock after scattering. This probability is given by the ratio of fluxes crossing the shock from one side to the other relative to the flux of particles crossing the shock from the other side to the first side. Figure 28 illustrates the case where we compare fluxes from right to left with those from left to right. In the downstream frame the shock moves to the left at speed $-u_d$ and all particles with velocity vectors in the shaded cone can catch up with the shock. All other particles pass the

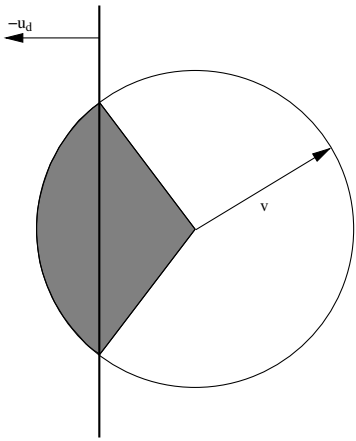


Fig. 28. Relative fluxes of shock-traversing particles.

shock from left to right. The normalized fluxes right to left and left to right are hence

$$\frac{|\int_{-v}^{-u} dv_x(u+v_x)|}{\int_{-v}^{v} dv_x} = \frac{(v-u)^2}{4v}, \tag{89}$$

$$\frac{|\int_{-u}^{v} dv_x(u+v_x)|}{\int_{-v}^{v} dv_x} = \frac{(v+u)^2}{4v}. \tag{90}$$

The probability to be reflected back is given by the ratio of the two expressions,

$$\mathcal{P}_i = \frac{(v_i-u)^2}{(v_i+u)^2} = \left(\frac{1-u/v_i}{1+u/v_i}\right)^2 \approx 1 - \frac{4u}{v_i} \quad \text{for } u \ll v_i. \tag{91}$$

We can now combine all ingredients to find the exponent of the spectrum, Γ, according to (78)

$$\begin{aligned} \Gamma(N) &= \frac{4u_d \sum_{i=1}^{N} \frac{1}{v_i}}{\frac{4}{3}(u_u - u_d)\sum_{i=1}^{N} \frac{1}{v_i}} \\ &= \frac{3u_d}{u_u - u_d} = \frac{3}{r-1}, \end{aligned} \tag{92}$$

where $r = u_u/u_d$ is the shock compression ratio. Hence the spectral index is given by the compression ratio. Inspite of the fact that both the incremental momentum gain and the single reflection probability depend of N and p, the exponent Γ is independent of these quantities. Because $r < 4$ for non-relativistic shocks, $\Gamma \geq 1$ and because most shocks are strong ($r \sim 4$), Γ lies near unity, in agreement with observations.

5.3 Transport Processes and Cosmic Ray Modulation

In the context of particle acceleration, the concept of diffusion is important. This is a very simple form of a more general concept called transport. Examples of transport equations include the ordinary diffusion equation, or the convection-diffusion equation, but also more complicated equations that describe energetic particles in the heliosphere. We have already found that particles are scattered at magnetic field irregularities, this can also be described as a transport process. The stationary one-dimensional convection-diffusion equation

$$\partial_r (v(r) I(r)) = \partial_r \left(\kappa(r) \partial_r I(r)\right), \tag{93}$$

is an extremely simple model of this situation. κ is the diffusion tensor, which is simplified here as a diffusion constant. Let us assume for simplicity that $\partial_r v(r) = 0$ and $\partial_r \kappa(r) = 0$. Then we can rewrite (93) as

$$\left(\partial_r - \frac{v - \partial_r \kappa}{\kappa}\right)(\partial_r - 0)\, I = 0, \tag{94}$$

which has the solution

$$I(r) = C_1 + C_2 e^{\frac{v - \partial_r \kappa}{\kappa} r}. \tag{95}$$

This behavior, shown in Fig. 29, describes the penetration of galactic cosmic rays into the heliosphere in a barely legal oversimplification. For instance it neglects completely the drifts which we considered earlier on in this section. Galactic cosmic rays penetrate the heliosphere, but are scattered by magnetic irregularities, especially the large barriers built up during solar activity maximum periods, global merged interaction regions (GMIRs), discussed in Sect. 4.5. This demonstrates why modulation in the heliosphere is not linear, as expected based on a non-convective diffusion equation, but exponential.

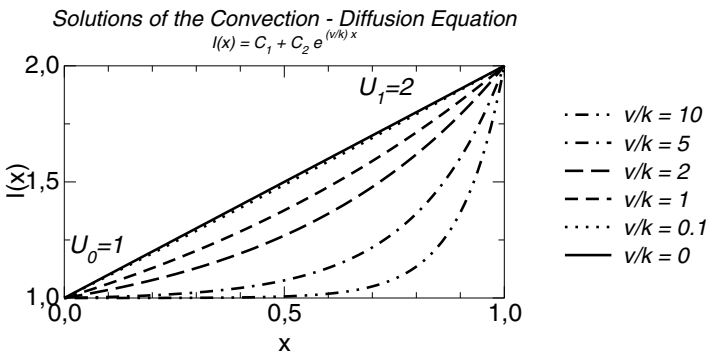


Fig. 29. Solutions of the one-dimensional diffusion-convection equation. For small convection speeds, the solution of the ordinary diffusion equation is recovered.

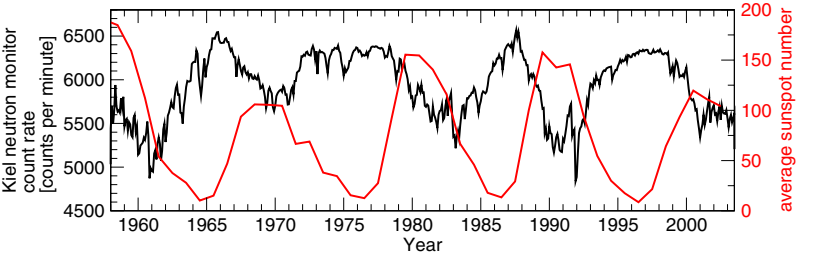


Fig. 30. Modulation of galactic cosmic rays. The count rate of neutron monitors (here the Kiel neutron monitor, providing one of the longest data sets) given with the left-hand scale is roughly anti-correlated with sunspot number, the smooth line, given on the right-hand scale.

For small v the usual linear behavior is recovered, as can be seen by explicitly calculating the integration constants C_1 and C_2.

Measurements of cosmic ray intensities can be performed at ground level on Earth with neutron monitors. For example, the Kiel neutron monitor has been measuring cosmic ray intensities for four sunspot cycles or two magnetic cycles, since 1957. Figure 30 shows the time series. Cosmic ray intensities vary by some tens of percents over the course of an activity cycle, but not in the same way in even and odd cycles. The reason for this difference lies in drifts in the heliosphere. Especially ∇B drift is important for galactic cosmic when they penetrate into the heliosphere along the heliospheric current sheet. Because this drift depends on the polarity of the fields, cosmic ray ions (mainly protons) penetrate along the current sheet more easily during one cycle, than during the next, this is partially responsible for the different shapes of the intensity profiles during even and odd activity cycles. Because CMEs appear to carry with them the current sheet, they are an important ingredient in the modulation of galactic cosmic rays.

6 Activity of the Sun in the Past and Comparison with Other Stars

6.1 Solar Rotation in Time

Because of the dearth of information about solar activity in the distant past we must resort to comparisons with other stars. We would expect solar and stellar activity to be somehow related to the generation of the magnetic field by the solar or stellar dynamo. This, in turn, is probably related to the rotation rate which must have been higher in the past. We will consider the slow-down of the Sun due to the loss of angular momentum to the escaping solar wind and compare this theoretical expectation with observations of other stars of differing ages. We follow the derivation given by Bochsler (1992) [5].

Solar wind particles leaving the Sun carry with them angular momentum, L. As they cross the Alfvén radius, r_A, they decouple from the Sun and remove this angular momentum at a rate

$$\dot{L} = -\beta \dot{m} r_A^2 \omega. \tag{96}$$

Here β is a dimensionless factor describing the geometry of the problem and r_A is the Alfvén radius, i. e. the radius where $v_{\rm sw} = v_A = B_A/\sqrt{\mu_0 \rho}$. Since solar wind outflow time is fast compared to the age of the Sun, this reduces the solar angular rotation

$$\dot{L} = \alpha m \dot{\omega} r_\odot^2 \tag{97}$$

where $\alpha = 2/5$ for a homogeneous sphere. Equating (96) and (97) we can find the spin-down rate of the Sun:

$$\frac{\dot{\omega}}{\omega} = -\frac{\beta \dot{m} r_A^2}{\alpha m r_\odot^2} \tag{98}$$

Mass loss is given by $\dot{m} = \rho v_A 4\pi R_A^2$ and we relate $B_A = B(r_A)$ to $B_\odot$ via $B_A R_A^2 = B_\odot R_\odot^2$. Thus

$$\begin{aligned}
\frac{\dot{\omega}}{\omega} &= -\frac{\beta}{\alpha} \frac{\rho v_a 4\pi R_A^2}{m} \frac{B_\odot}{B_A} \\
&= -\frac{\beta}{\alpha} \frac{\rho B_A}{\sqrt{\mu_0 \rho}} \frac{4\pi R_A^2}{m} \frac{B_\odot}{B_A} \\
&= -\frac{\beta}{\alpha} \frac{\rho v_a 4\pi R_A^2}{m} \frac{B_\odot}{B_A} \\
&= -\frac{\beta}{\alpha} \frac{\rho B_A}{\sqrt{\mu_0 \rho}} \frac{4\pi}{m} R_A^2 \frac{B_\odot}{B_A} \\
&= -\frac{\beta}{\alpha} \frac{\mu_0}{\mu_0} \frac{\rho}{B_A} \frac{1}{\sqrt{\mu_0 \rho}} \frac{4\pi}{m} B_\odot^2 R_\odot^2 \frac{B_A}{B_A} \\
&= -\frac{\beta}{\alpha} \frac{\mu_0 \rho}{B_A^2} \frac{B_A}{\sqrt{\mu_0 \rho}} \frac{4\pi}{\mu_0 m} B_\odot^2 R_\odot^2, \\
\frac{\dot{\omega}}{\omega} &= -\frac{\beta}{\alpha} \frac{4\pi B_\odot^2 R_\odot^2}{v_A m \mu_0},
\end{aligned}$$

i. e. the entire complicated behavior is hidden in v_A and, possibly, α and β. Note that $\dot{m}$ has vanished, it has been absorbed in v_A.

Assuming that

$$B_\odot(t) = B_\odot(\tau)[\omega(t)/\omega(\tau)], \tag{99}$$

we have $v_A(t) = v_A(\tau)[\omega(t)/\omega(\tau)]^{1/2}$, where τ is today. Inserting

$$\begin{aligned}
\frac{\dot{\omega}}{\omega} &= -\frac{\beta}{\alpha}\frac{R_\odot^2 B_\odot^2(\tau)4\pi}{m v_A(\tau)\mu_0}\left(\frac{\omega(t)}{\omega(\tau)}\right)^{3/2}, \\
\dot{\omega} &= -\frac{A}{\omega^{3/2}(\tau)}\omega^{5/2}(t), \\
\frac{d\omega}{dt} &= -\frac{A}{\omega^{3/2}(\tau)}\omega^{5/2}(t), \\
d\omega\omega^{-5/2}(t) &= -\frac{A}{\omega^{3/2}(\tau)}dt, \\
-\frac{2}{3}\omega^{-3/2}(t) &= -\frac{A}{\omega^{3/2}(\tau)}t + C,
\end{aligned}$$

The constant of integration is determined by the final condition that $\omega(t = \tau) = \omega(\tau)$,

$$\begin{aligned}
\omega^{-3/2}(t = \tau) &= \frac{3}{2}A\omega^{-3/2}(\tau)\tau + C, \\
C &= \omega^{-3/2}(\tau)\left(1 - \frac{3}{2}A\tau\right),
\end{aligned}$$

and the final solution for the temporal behavior of stellar rotation is:

$$\begin{aligned}
\omega^{-3/2}(t) &= \frac{3}{2}A\omega^{-3/2}(\tau)t + \omega^{-3/2}(\tau)\left(1 - \frac{3}{2}A\tau\right), \\
&= \omega^{-3/2}(\tau)\left(1 + \frac{3}{2}A(t - \tau\right), \\
\omega(t) &= \frac{\omega(\tau)}{\left(1 + \frac{3}{2}A\,(t - \tau)\right)^{2/3}}. \qquad (100)
\end{aligned}$$

This is similar to the behavior originally observed by Skumanich (1972) [43]. He found that empirically, $\omega(t) \sim t^{-1/2}$ and this relation or similar power-law rotation-rate evolution has often been called "Skumanich-rule". The fact that we did not derive a $t^{-1/2}$ dependence, but rather a $t^{-2/3}$ power law, is simply due to our choice of parameterization of the Alfvén speed, v_A. A wealth of data on stellar rotation rates exist today that can be used to infer solar activity in the past. We will compare the change of solar rotation rate with the rotation rates of stars of different ages in the next section.

The faster solar rotation implies stronger magnetic fields on the Sun in the past. If we assume that the solar wind flux is related to the available magnetic energy, $\varphi_{\rm SW} \propto B^2$, then we can crudely estimate the solar wind flux in the past. In (99) we assumed that B was proportional to ω, implying that $\varphi_{\rm SW} \propto \omega^2$. Thus we can find the enhancement factor of solar wind flux

in the past relative to its present value

$$\frac{\varphi_{\mathrm{SW,\ past}}}{\varphi_{\mathrm{SW,\ today}}} = \frac{\int_1^{4.57} dt\ \left(\omega(t)/\omega(\tau)\right)^2}{\int_1^{4.57} dt} = 2.44. \tag{101}$$

The enhancement by about a factor of 2 is the same as that inferred by Geiss (1973) [13] from measurements of lunar soil Kr and Xe.

6.2 Inferences from Stellar Activity

From analysis of historic records of sunspots we know that the Sun has other cycles superimposed on the 11 year solar activity cycle. These historic records are complemented by geologic records dating back several millennia. Together, these records indicate that the Sun's activity drops to levels below that of solar activity minimum, for instance during the Maunder minimum. Obviously, the heliosphere will behave quite differently during such time periods. For instance, analysis of the Be production in terrestrial archives during the Maunder minimum has led Beer et al. (1998) [4] to conclude that the Sun could modulate the GCR during that time period. This probably shows that CIRs are also effective modulating agents for the GCR (see e. g. McKibben et al., 1999 [30] for a review). However, for our study, and for many others, geologic and historic records are much too short, we are interested in solar activity spanning hundreds of millions to billions of years.

Fortunately, solar activity also manifests itself in ways that allows us to compare it with activity of other stars. Leighton (1959) [24] showed that areas with more concentrated magnetic field emit two lines of singly charged calcium more strongly than areas with average field strength. The Ca II lines are labelled H (396.8 nm) and K (393.4 nm). The emission intensity of the Ca II H and K lines corresponds to the product of the area of the emitting region times the strength of the magnetic field in it (Skumanich et al., 1975 [44]). Thus this emission is a sensitive measure of solar activity.

Emission of Ca H and K can also be measured for other stars and in this way, solar activity can be compared with that of other stars. Ca II H and K emission has been measured for lower main-sequence stars since 1957 (Wilson and Bappu, 1957 [58]). Most information comes from a program started in 1966 and is called the "HK Project" (Wilson, 1968 [56]; Wilson, 1978 [57]). Such investigations have confirmed the basic Skumanich law (Skumanich, 1972 [43]) that young stars rotate faster and are more active than old ones. From these investigations we also know that the very young Sun should have been magnetically more active (since faster rotation most probably implies stronger dynamo actions) and should have had no Maunder-like activity minima. Comparison of stars of similar ages as the Sun shows that most have activity cycles similar to that of the Sun. Maunder-type activity minima also seem to occur. At intermediate ages, the Sun probably had intermediate rotation speeds and activity, occasionally with smooth activity cycles, similar

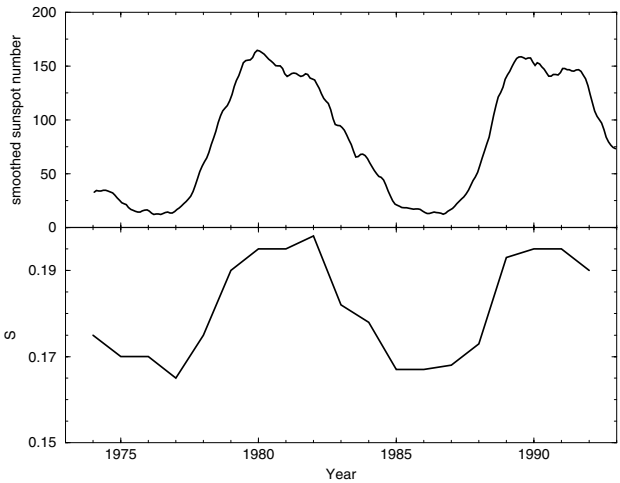


Fig. 31. Monthly averaged sunspot number (top) and chromospheric activity index S bottom plotted versus time.

to the present-day Sun (Baliunas et al., 1995 [2]). These authors measured chromospheric activity periods in just over 50 stars and the remainder of the investigated 111 stars had either long-term trends, variable activity without a clear period, or show flat activity curves. We have plotted the distribution of an activity measure, S, from monthly averages for a sample of 74 stars reported by Baliunas and Jastrow (1990) [1] in Fig. 32. The lower panel shows the number of occurrences of monthly average values of S for solar-type stars. The upper panel shows the cumulative sum of the numbers in the lower panel. The solar value for S ranges between 0.164 and 0.178 with an average[5] of 0.171. The distribution in the lower panel is bimodal with a relatively well defined peak at low activity and a broad distribution at higher activity. The broad distribution is due to a convolution of two contributions. While the sample consists of solar-type stars they do not necessarily all have the same average activity. Second, their activity cycles spread out the distribution of average activity. The narrow peak at low activity has been interpreted as indicative of Maunder-type activity minima in a substantial number of stars. The fraction of stars in this sample currently in such activity minima can be read off the top panel, it is about 25%. Assuming that the Sun is not unusual and that the sample indeed consists of solar-type stars, we can infer that the Sun has spent on average about one quarter of its past in Maunder-type activity minima. In the more distant past (when the Sun was about 1 Gy old),

[5] The reason for the discrepancy of the values for the solar S quoted here and in the Baliunas and Jastrow (1990) [1] paper and those given in Fig. 31 is not obvious to us. The data are from the same authors. It is possible that the discrepancy is due to a recalibration of the S index that took place during that time period.

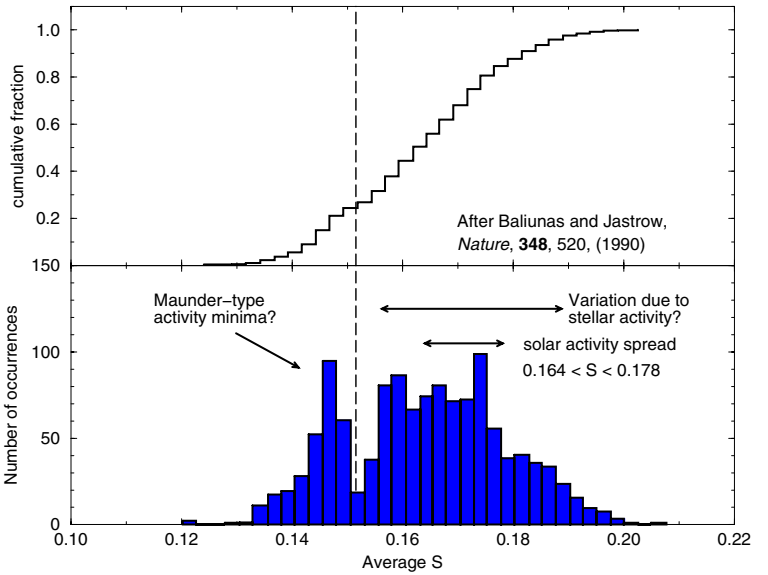


Fig. 32. Lower panel: Number of occurrences of monthly average values of S for solar-type stars versus average S. Data from Baliunas and Jastrow (1990) [1]. Top panel: Cumulative sum of lower panel.

such minima probably did not occur, and there may have been intermediate active/minimum ratios during the intermediate part of the life of the Sun. We need to note that these statements are presently all of a statistical nature. None of the stars that are assumed to be undergoing an activity minimum has been observed to exit from it yet, nor has an active star been seen to enter an activity minimum. Before the interpretation advanced by Baliunas and Jastrow (1990) [1] can be accepted as an established fact at least one star should be observed to make the transition from one activity state to another. However, intuitively, their interpretation is appealing.

Observations of short-term time series of chromospheric activity has allowed to determine the rotation periods of a number of stars (Noyes et al., 1984 [34]). Soderblom et al. (1991) [46] have found a relation of chromospheric activity with age

$$\log(\text{age}) = -1.5 \log(R'_{HK}) + 2.25, \tag{102}$$

where R'_{HK} is a chromospheric activity index that has been corrected for color effects. Thus we can compare our expression for the solar rotation period (100) with the rotation periods of other stars of different ages. Because the stellar rotation period is not inferred from R'_{HK}, but has been determined by analyzing time series of this quantity, we are not arguing in a circle, the two measurements are not linked in a logical way. Rotation periods could, in principle, have been determined in different ways. In Fig. 33 we have plotted

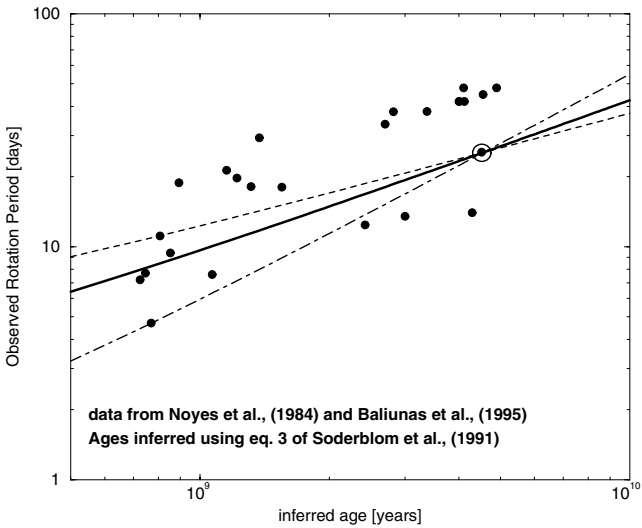


Fig. 33. Stellar rotation versus age inferred from (102). The solid curve shows our expectation based on (100) while the dashed curve shows the Skumanich law and the dash-dotted curve is for an exponent of unity in (100). Data are from Noyes et al. (1984) [34].

observed stellar rotation periods versus for a number of stars (Noyes et al., 1984 [34]) versus their age inferred according to (102).

The circled symbol is the Sun and the solid curve is described by (100), the expression for the time dependence of solar rotation. We observe quite some scatter round the curve which is due to the spread in initial angular momentum of the stars and uncertainties in their ages. Equation (100) appears to describe the trend for deceleration with time remarkably well. The two other curves are for different exponents in the denominator of (100). The dashed curve with slower rotation in the past (the upper curve for young stars) corresponds to a Skumanich scaling (Skumanich, 1972 [43]) with exponent 1/2, while the dash-dotted curve for faster rotation in the past (the lower curve for young stars) shows the behavior for an exponent of unity. Obviously, the solid curve does the best job of the three curves.

We observe that the exponent 2/3 in (100) is the same as the time exponent in (102). This is not by construction and probably not a coincidence either. It implies that the relationship between chromospheric activity and rotation rate is linear. Because dynamo action which is responsible for generating the solar (and stellar) magnetic field is intimately connected to rotation (differential rotation, to be precise) we do expect a causal relationship. Given the complexity of dynamo physics, the simplicity of this empirical relation is indeed quite remarkable.

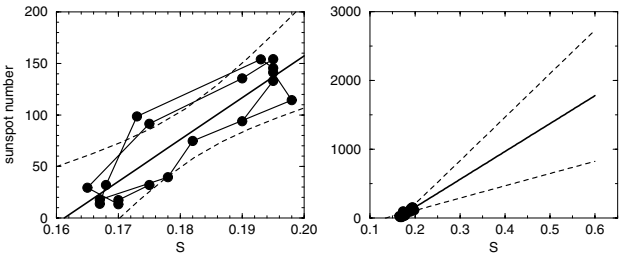


Fig. 34. Sunspot number versus chromospheric activity index S for present-day solar activity (left panel) and an extrapolation to possible activity levels in the past (right panel). Dashed lines show the 95% confidence levels for the expectation value for sunspot number based on the fit in the left panel.

We can now proceed to estimate solar activity in the past. We relate sunspot number to chromospheric activity index S in Fig. 34. The left-hand panel shows yearly averaged sunspot numbers (derived from smoothed monthly averages) plotted versus solar chromospheric activity index S. The two quantities are well correlated, the solid line shows the best linear fit and the two dashed lines the 95% confidence level limits in the expectation value for sun spot number for a given value of S. Another solid curve connects subsequent data points ranging from 1974 to 1992. The data are the same as those used in Fig. 31. The right-hand panel shows an extrapolation to large values for S of the relation found in the left-hand panel. From Fig. 33 we can read off that the Sun rotated about 2.5 times faster when it was about 1 Gy old. Because of the linear relation between activity and rotation we know that solar activity must have been about 2.5 times higher at that time. Thus a maximum value for S would lie around $S_{\max} \approx 0.5$. This corresponds to an expected yearly sunspot number of about 1400. As attractive as it may appear, this inferred strong enhancement of solar activity in the past is probably misleading. The fit in Fig. 34 would require negative sunspot numbers for $S \lesssim 0.16$. This is certainly not true. Assuming that the peak for low values of S in Fig. 32 is indeed due to stars undergoing Maunder-type activity minima, we see that we need to have a small, non-negative number of sunspots for values of $S \sim 0.14$. To estimate solar activity in the past we now assume that $S_{\min} \sim 0.145$ corresponds to the minimal possible activity, and that the difference $S - S_{\min}$ is time dependent. This gives us a maximum value for S, $S_{\max} \sim 0.14 + S_{\min} \sim 0.285$. This would imply a yearly average of about 500 sunspots when the Sun was 1 Gy old. If the minimum activity index were $S \sim 0.16$ this number would shrink to about 320. Simply scaling sunspot number by a factor of 2.5 for solar rotation would imply a sunspot number of about 380 for the early Sun. In the coming discussion we will assume a sunspot number of 400 when the Sun was 1 Gy old.

From Fig. 21 we infer a daily coronal mass ejection rate of about 6 CMEs per day at the Sun. From Fig. 22 we see that, for average activity, only about 6% of all CMEs result in sudden commencements which in turn indicate a CME-driven shock. Thus we could reckon with a maximum incidence rate of about 10 CME-driven interplanetary shocks per month 3.5 Gy ago. This is not that much more than the present-day CME incidence rate.

Acknowledgments

I wish to thank the organizers for their work and DPG for their support. This work has profited from work by numerous other authors which could not always be identified in the text. Among them are the books by Kallenrode (1998) [21], Rossi and Olbert (2003) [37], and Parker (1979) [35], but also of Chen (1984) [7], Stone and Tsurutani (1985) [50], ...

References

1. Baliunas, S. and R. Jastrow: *Nature*, **348**, (1990), 520 – 523.
2. Baliunas, S. L., R. A. Donahue, W. H. Soon, J. H. Horne, J. Frazer, L. Woodard-Ecklund, M. Bradford, L. M. Rao, O. C. Wilson, Q. Zhang, *et al.*: *Astrophys. J.*, **438**, (1995), 269 – 287.
3. Balogh, A., V. Bothmer, N. Crooker, R. Forsyth, G. Gloeckler, A. Hewish, M. Hilchenbach, R. Kallenbach, B. Klecker, J. Linker, *et al.*: *Space Sci. Rev.*, **89**, (1999), 141 – 178.
4. Beer, J., S. Tobias, and N. Weiss: *Sol. Phys.*, **181**, (1998), 237 – 249.
5. Bochsler, P.: Physikalisches Institut, University of Bern (1992). Skriptum zur Vorlesung.
6. Bothmer, V. and D. M. Rust: In , edited by N. Crooker, J. A. Joselyn, and J. Feynman, Geophysical Monograph Series. American Geophysical Union, Washington DC, USA (1997). 139 – 146.
7. Chen, F.F., Plasma Physics and Controlled Fusion, Plenum Press, New York, 1984.
8. Crooker, N. U., J. T. Gosling, and S. W. Kahler: *J. Geophys. Res.*, **103**, (1998), 301 – 306.
9. Debrunner, H., E. Flueckiger, E. L. Chupp, and D. J. Forrest: *International Cosmic Ray Conference, 18th, Bangalore, India, August 22-September 3, 1983, Conference Papers. Volume 4 (A85-22801 09-93). Bombay, Tata Institute of Fundamental Research*, **4**, (1983), 75–78.
10. Dulk, G. A.: *Ann. Rev. Astron. Astrophys.*, **23**, (1985), 169 – 224.
11. Eddy, J. A.: *Science*, **192**, (1976), 1189 – 1202.
12. Fermi, E.: *Phys. Rev.*, **75**, (1949), 1169 – 1174.
13. Geiss, J.: volume 5 (1973). 3375 – 3398. Proceedings of 13th International Cosmic Ray Conference.

14. Gloeckler, G., L. A. Fisk, S. Hefti, N. Schwadron, T. Zurbuchen, F. M. Ipavich, J. Geiss, P. Bochsler, and R. Wimmer-Schweingruber: *Geophys. Res. Lett*, **26**, (1999), 157 – 160.
15. Gloeckler, G., J. Geiss, E. C. Roelof, L. A. Fisk, F. M. Ipavich, K. W. Ogilvie, L. J. Lanzerotti, R. von Steiger, and B. Wilken: *J. Geophys. Res.*, **99**, (1994), 17637 – 17643.
16. Golub, L. and J. M. Pasachoff: *The Solar Corona* Cambridge University Press, Cambridge, UK (1997).
17. Gosling, J. T., P. Riley, D. J. McComas, and V. J. Pizzo: *J. Geophys. Res.*, **103**, (1998), 1941 – 1954.
18. Henke, T., J. Woch, U. Mall, S. Livi, B. Wilken, R. Schwenn, G. Gloeckler, R. v. Steiger, R. J. Forsyth, and A. Balogh: *Geophys. Res. Lett.*, **25**, (1998), 3465 – 3468.
19. Hurford, G.J., Schwartz, R.A., Krucker, S., Lin, R.P., Smith, D.M., Vilmer, N., *Astrophys. J.*, **595**, (2003), 77 – 80.
20. Jones, F. C.: *Astrophys. J. Suppl. Ser.*, **90**, (1994), 561 – 665.
21. Kallenrode, M.-B., Space Physics, Springer, Berlin, 1998.
22. Kohl, J. L., R. Esser, L. D. Gardner, S. Habbal, P. S. Daigneau, E. F. Dennis, G. U. Nystrom, A. Panasyuk, J. C. Raymond, P. L. Smith, *et al.*: *Solar Phys.*, **162**, (1995), 313 – 356.
23. Lang, K. R.: *The Sun from Space* Springer, Berlin (2000).
24. Leighton, R. B.: *Astrophys. J.*, **130**, (1959), 366 – 380.
25. Lin, R.P., Krucker, S., Hurford, G.J., Smith, D.M., Hudson, H.S., et al., *Astrophys. J.*, **595**, (2003), 69 – 76.
26. Lindsay, G. M., C. T. Russell, J. G. Luhmann, and P. Gazis: *J. Geophys. Res.*, **99**, (1994), 11 – 17.
27. Low, B. C.: In , edited by N. Crooker, J. A. Joselyn, and J. Feynman. American Geophysical Union (1997). 39 – 47. Geophysical Monograph 99.
28. Maunder, E. W.: *Mon. Not. Royal Astron. Soc.*, **50**, (1890), 251.
29. Maunder, E. W.: *Knowledge*, **17**, (1894), 173.
30. McKibben, R. B., J. R. Jokipii, R. A. Burger, B. Heber, J. Kóta, F. B. McDonald, C. Paizis, M. Potgieter, and I. G. Richardson: *Space Sci. Rev.*, **89**, (1999), 307 – 326.
31. Neukomm, R. O. and P. Bochsler: *Astrophys. J.*, **465**, (1996), 462 – 472.
32. Nindos, A. and H. Zhang: *Astrophys. J.*, **573**, (2002), L133 – L136.
33. Nindos, A., J. Zhang, and H. Zhang: *Astrophys. J.*, **594**, (2003), 1033 – 1048.
34. Noyes, R. W., L. W. Hartmann, S. L. Baliunas, D. K. Duncan, and A. H. Vaughan: *Astrophys. J.*, **279**, (1984), 763 – 777.
35. Parker, E. N.: *Cosmical Magnetic Fields* Oxford University Press (1979).
36. Petschek, H. E.: In , edited by W. N. Ness. NASA SP-50 (1964). 425 – 439.
37. Rossi, B. Olbert, S., Introduction to the Physics of Space, McGraw-Hill, 1970

38. Schüssler, M.: In , edited by K. Scherer, H. Fichtner, B. Heber, and U. Mall, Lecture Notes in Physics, Springer, Berlin, Germany (2003), this volume.
39. Schwabe, H.: *astron. Nachr.*, **20, No.205**, 1843.
40. Schwadron, N. A., L. A. Fisk, and G. Gloeckler: *Geophys. Res. Lett.*, **23**, (1996), 2871 – 2874.
41. Schwenn, R., H. Rosenbauer, and K.-H. Mühlhäuser: *Geophys. Res. Lett.*, **7**, (1980), 201 – 204.
42. Shea, M. A. and D. F. Smart: *Solar Phys.*, **127**, (1990), 297 – 320.
43. Skumanich, A.: *Astrophys. J.*, **171**, (1972), 565 – 567.
44. Skumanich, A., C. Smythe, and E. N. Frazier: *Astrophys. J.*, **200**, (1975), 747 – 764.
45. Smith, E. J.: *Space Sci. Rev.*, **34**, (1983), 101 – 110.
46. Soderblom, D. R., D. K. Duncan, and D. R. H. Johnson: *Astrophys. J.*, **375**, (1991), 722 – 739.
47. Spörer, F. W. G.: *Vierteljahrsschrift Astron. Ges. Leipzig*, **22**, (1887), 323.
48. Spörer, F. W. G.: *Bull. Astron.*, **6**, (1889), 60.
49. St. Cyr, O. C., R. A. Howard, N. R. Sheeley Jr., S. P. Plunkett, D. j. Michels, S. E. Paswaters, M. J. Koomen, G. M. Simnett, B. j. Thompson, J. B. Gurman, *et al.*: *J. Geophys. Res.*, **105**, (2000), 18169 – 18185.
50. Stone, R.G., Tsurutani, B.T., Collisionless Shocks in the Heliosphere: A Tutorial, American Geophysical Union, 1985.
51. Webb, D. F. and R. A. Howard: *J. Geophys. Res.*, **99**, (1994), 4201 – 4220.
52. Widing, K. G. and U. Feldman: *Astrophys. J.*, **555**, (2001), 426 – 434.
53. Wieler, R.: *Space Sci. Rev.*, **85**, (1998), 303 – 314.
54. Wieler, R., H. Baur, and P. Signer: *Geochim. et Cosmoschim. Acta*, **50**, (1986), 1997 – 2017.
55. Wilhelm, K., W. Curdt, E. Marsch, U. Schüle, P. Lemaire, A. Gabriel, J. Vial, M. Grewing, M. C. E. Huber, S. D. Jordan, *et al.*: *Solar Phys.*, **162**, (1995), 189 – 231.
56. Wilson, O. C.: *Astrophys. J.*, **153**, (1968), 221 – 234.
57. Wilson, O. C.: *Astrophys. J.*, **226**, (1978), 379 – 396.
58. Wilson, O. C. and M. K. V. Bappu: *Astrophys. J.*, **125**, (1957), 661 – 683.
59. Wimmer-Schweingruber, R. F.: In , edited by M. Velli. American Institute of Physics, Melville, NY, USA (2003). Accepted for publication.
60. Wimmer-Schweingruber, R.F., Bochsler, P., Lunar Soils: A Long-Term Archive for the Galactic Environment of the Heliosphere?, in: AIP conference proceedings, 399 – 404, 2001.
61. Wimmer-Schweingruber, R. F., O. Kern, and D. C. Hamilton: *Geophys. Res. Lett.*, **26**, (1999), 3541 – 3544.
62. Wolf, R.: *Astron. Mitt. Zürich*, **14**

The Magnetosphere

Antonius Otto

Geophysical Institute, University Alaska Fairbanks, AK 99775-7320, USA

Abstract. First, a brief history of magnetospheric research and a description of the basic structure of the magnetosphere is given. Following are detailed discussions of the major magnetospheric structures, i.e. the bow shock, the magnetosheath, the magnetopause, and the magnetotail. Finally, the role of the inner magnetosphere for geomagnetic storms, which are important mainfestations of space weather, is explained.

1 Introduction

1.1 History

The *aurora* has been known since ancient times and is the most obvious and the beautiful manifestation of 'space weather' or the interaction of the Earth's magnetosphere with the solar wind. The name after a roman goddess has been introduced by Galileo Galilei (1564-1642). The second early observations of magnetospheric processes are *magnetic field measurements* which identified the geomagnetic field as mostly dipolar (William Gilbert, chief physician to Queen Elizabeth, 1600) with significant temporal magnetic perturbations (George Graham, 1722) which are now addressed as geomagnetic activity. In the eighteenth century a connection between the aurora and sunspot numbers on one side and between the aurora and magnetic activity on the other side has been suspected (Russel, 1995a [83]).

On September 1, 1859 Richard Carrington observed a great flare (duration of only a minute) and almost at the same time magnetic perturbations are recorded. 18 hours later one of the largest magnetic storms ever recorded developed. Today we know that the immediate response is caused by the increase of ionospheric conductance due to the UV radiation and the 18 hour delay is caused by the travel time of the shock in the solar wind.

In the following decades progress in experimental physics (gas discharges), observational methods (photography), and theoretical physics (Maxwell's equations) provided the the foundation for a much better understanding for the causes of the aurora and magnetic activity. In particular Kristian Birkeland (1867-1917) contributed to the science through experiments with cathode rays and his famous terella experiments which led him to conclude that the aurora is caused by energetic particles ejected from sunspots and guided by the magnetic field (Egeland, 1984 [28]). He categorizes magnetic

A. Otto, The Magnetosphere, Lect. Notes Phys. **656**, 133–192 (2005)
http://www.springerlink.com/

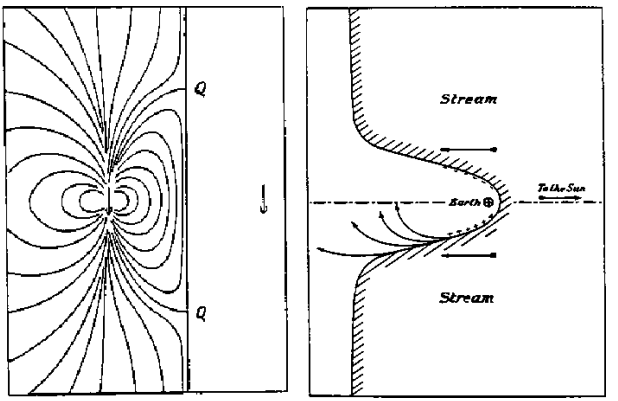


Fig. 1. Sketch of the interaction of the magnetosphere with and infinitely conducting solar wind (Chapman and Bartels, 1940 [22]).

perturbations (first finding of geomagnetic storms and auroral substorms) and identifies ionospheric currents as responsible for magnetic perturbations. His most famous contribution is the conclusion that large magnetic field-aligned currents must be present in the Aurora (Dessler, 1984 [26]).

However, progress is not always in a straight line. For instance Lord Kelvin (1892, president of the royal astronomical society) suggested "that the supposed connection between magnetic storms and sunspots is unreal". While the idea of strong field-aligned currents remained almost forgotten Sydney Chapman dominated the field in the first half of the twentieth century with theories on geomagnetic disturbances based on (equivalent) currents driven by atmospheric motion. The Chapman-Ferraro theory of currents and the compression of the magnetosphere is the first idea of a magnetopause (Chapman and Ferraro, 1930 [23]), i.e., a boundary between the magnetosphere and the ambient medium (Fig. 1). In 1925 Appleton detects the previously expected ionosphere using radio waves.

In the second half of the twentieth century magnetospheric physics moves into the realm of modern plasma physics. Biermann (1951) [7] predicts a continuous solar wind consisting of charged particles and Hannes Alfvén suggests that the solar wind particles are magnetized (Alfvén, 1957 [2]). Alfvén also revives Birkeland's idea of field-aligned currents and one of his most important contributions is the development of plasma fluid equations which are used by Parker for a quantitative theory of the solar wind (Parker, 1958 [71]).

Until the early 60ths the magnetosphere has been assumed to be closed meaning the Earth's magnetic field is entirely contained in a closed cavity called the magnetosphere. This idea was motivated by the consideration that there are no magnetic monopoles, i.e., all magnetic field must be closed thus providing an easier approach for mathematical models. However, the first theories of magnetic reconnection by Sweet (1958) [108] and Parker (1957) [70]

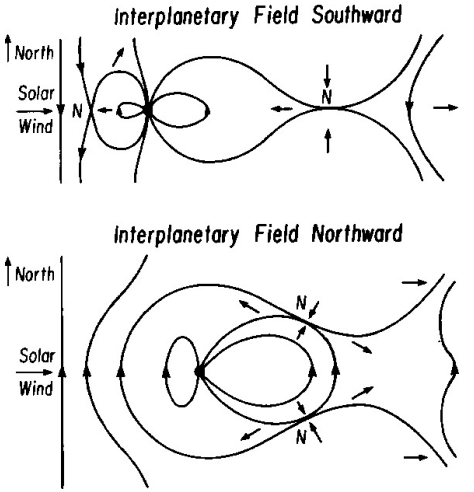


Fig. 2. Magnetic reconnection after Dungey (1961) [27].

led Dungey (1961) [27] to suggest the concept of an open magnetosphere (Fig. 2). In this theory geomagnetic field actually is connected to the solar wind. This resolved a major problem of prior models. It was very clear that charged particles from the solar wind could not penetrate deep enough into a closed magnetosphere to cause the Aurora. Petschek (1964) [76] develops the first theory of fast reconnection. The concept of viscous interaction between the solar wind and the magnetosphere is suggested by Axford and Hines (1961) [5]. The final major building block in this early era of magnetospheric physics is the description and theory of the auroral substorm by Akasofu (1964) [1].

During the past decades these ideas formed the framework for for modern space physics. Advances in observations, in particular in situ satellite measurements, but also optical and Radar techniques have generated much insight into magnetospheric plasma processes. Further development of theoretical models and during the past two decades numerical simulations have provided physical mechanisms and quantitative models of many fundamental plasma processes.

1.2 Basic Structure of the Magnetosphere

The magnetosphere is a large plasma cavity generated by the Earth's magnetic field and the solar wind plasma (Fig. 3). The streaming solar wind compresses the dayside portion of the Earth's field and generates a tail which is many hundreds of Earth radii long. The basic mechanism for the formation of the magnetosphere is fairly simple: The Earth's magnetic dipole field is exposed to a stream of charged particles. The entire magnetosphere is subject

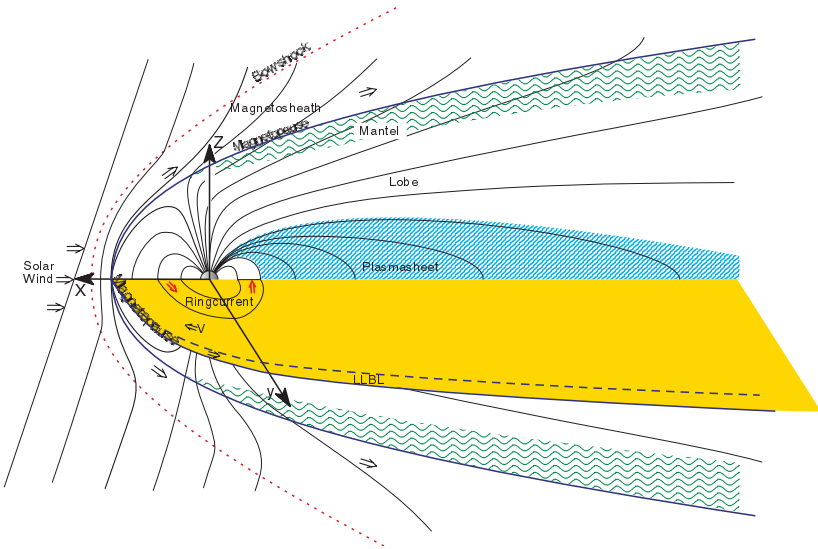


Fig. 3. Sketch of the structure of the magnetosphere.

to only two boundary conditions: (a) The boundary between the magnetosphere and the streaming solar wind and (b) the boundary of the magnetosphere and the neutral atmosphere. The basic structural elements of the magnetosphere are

- The Bow Shock and the Magnetosheath
- The Magnetopause
- The Magnetotail - Magnetic Substorms
- The Inner Magnetosphere - Magnetic Storms

The Bow Shock and the Magnetosheath are not actually part of the magnetosphere proper but form an outer layer embedding the magnetosphere. The solar wind plasma travels usually at super-fast speeds (faster than the fast mode speed) relative to the magnetosphere. Therefore a standing shock wave forms around the magnetosphere just as in front of an aircraft traveling at supersonic speeds. The bow shock is the shock in front of the magnetosphere and the magnetosheath is the shocked solar wind plasma. Therefore it is not directly the solar wind plasma which constitutes the boundary of the magnetosphere but the strongly heated and compressed plasma behind the bow shock .

The Magnetopause is the actual boundary between the shocked solar wind and the magnetospheric plasma. However, the magnetosphere is not closed in terms of the magnetic field but there is considerable magnetic flux crossing the magnetopause. Thus it is not easy to define this boundary in precise mathematical terms. The boundary permits a certain amount of solar wind plasma entry. This entry is easier along magnetic field lines. The

magnetopause is a highly important region because the physical processes at this boundary control the entry of plasma, momentum, energy into the magnetosphere.

The Magnetotail is the long tail-like extension of the magnetosphere on night side. Since the magnetic field points toward the Earth in the northern lobe and away in the southern lobe there is a current in the westward direction. Because of this structure there is considerable energy stored in the magnetic field in the magnetotail. During magnetically quiet times convection is typically slow and the energy in the plasma flow is only a tiny fraction of the overall energy density.

The magnetotail plays a particularly important role in so-called magnetospheric substorms. A substorm is characterized by a specific large scale auroral intensifications and corresponding magnetic perturbations at high latitudes typically close to magnetic midnight. Characteristic are also fast flows in the magnetotail, plasma ejection in the tailward direction, release of energy stored in the lobe magnetic field, energetic particle injections at geosynchronous distances, and strong intensifications of field-aligned current systems. A substorm consists of the growth phase, the (auroral) expansion phase, and the recovery phase. Substorms are clearly related to periods of southward interplanetary magnetic field (IMF) which leads to reconnection on the dayside, transport of magnetic flux from the dayside to the tail, and storage of magnetic energy in the tail during the growth phase which subsequently released in the expansion phase. Substorms are an important aspect for space weather because of large perturbations in magnetic and electric fields and because of the generation of high energy particle populations.

The terminology of magnetic storms and substorms is misleading in that substorms are not small storms. Rather a storm can consist of several substorms but also of quiet periods.

The Inner Magnetosphere is different from most of the magnetosphere in that the magnetic field is mostly dipolar and perturbations of the field are small compared to the average dipole field. However, there can still be large amounts of energy stored in this region in particular during magnetic storms. During such times the ring current (current due to gradient curvature drifts of charged particles around the Earth) intensifies strongly and is responsible for strong magnetic perturbations at low geomagnetic latitudes on the Earth.

A storm is a large and long duration (few days) perturbation of the magnetosphere which leads to a strong compression and a contraction of the magnetosphere. Characteristic is a strong amplification of the ring current and the associated magnetic field measured at equatorial latitudes. Aurora is typically visible at much lower latitudes. Storms are associated with larger solar flares and/or coronal mass ejections (CME's) and storm durations are many hours.

1.3 Other Remarks

The terrestrial magnetosphere is not unique. All planets with a sufficient intrinsic magnetic field exposed to streaming plasma have magnetospheres (Fig. 4). Similarly astrophysical objects such as pulsars or galaxies with their own magnetic field have magnetosphere. The size of planetary magnetospheres depends strongly on the intrinsic field strength. However, the global structure can vary considerably depending on the orientation of the magnetic field relative to the solar wind plasma stream.

Thus the study of the terrestrial magnetic environment and space weather has important implications not only for our planetary system but for astrophysics in general. It should be noted that more than 99% of the visible universe is in the plasma state.

Plasma properties vary strongly within the Earth's magnetosphere (Fig. 5). Plasma densities vary between some 10^6 particles per cm^3 in the ionosphere down to about 10^{-2} particles per cm^3 in the lobes of the magnetotail. Similarly magnetic field strength varies from about 10^{-9} Tesla in the tail plasma sheet up to about 10^{-4} Tesla in the ionosphere. Similarly the fundamental plasma parameters such as the Debye length λ_D and the electron plasma frequency ω_{pe}

$$\lambda_D = \left(\frac{\epsilon_0 k_B T_e}{n_e e^2}\right)^{1/2} \qquad \omega_{pe} = \left(\frac{n_e e^2}{m_e \epsilon_0}\right)^{1/2}$$

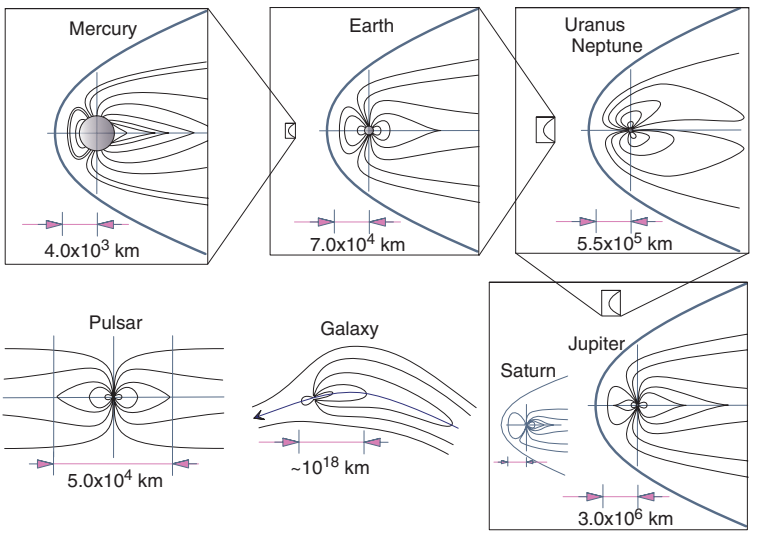


Fig. 4. Comparison of different planetary magnetospheres.

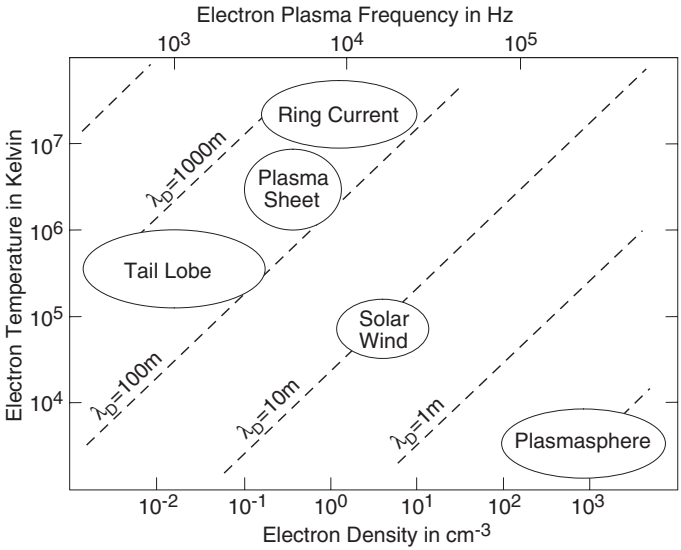


Fig. 5. Plasma parameters in the Earth's magnetosphere.

vary over several orders of magnitude. Here $\epsilon_0 = 8.85{\cdot}10^{-12}\mathrm{Fm}^{-1}$, Boltzmann constant $k_B = 1.38 \cdot 10^{-23}\mathrm{JK}^{-1}$, elementary charge $e = 1.6 \cdot 10^{-19}\mathrm{C}$, and $m_e = 9.11 \cdot 10^{-31}\mathrm{kg}$.

The plasma in most of the magnetosphere is highly ionized. In fact there is only a very thin skin of a few hundred km thickness in the ionosphere where the plasma is only partially ionized and dominated by collisions mostly with the neutral atmosphere. In the vast body of the magnetosphere the plasma can be treated as almost collisionless. While this simplifies the plasma dynamics considerably this also exposes fundamental problems in particular with processes of plasma entry into the magnetosphere and magnetic reconnection in the magnetosphere.

Magnetospheric plasma processes can be treated in different plasma approximations where any approximation should be appropriate in terms of the assumptions to its particular application. The two basic approximations are the so-called Vlasov equations which represent a kinetic treatment of a collsionless plasma and the one fluid or magnetohydrodynamic equations (e.g. Kivelson and Russel, 1995 [40], Baumjohann and Treumann, 1997 [6]). In the following we will mostly refer to the magnetohydrodynamic (MHD) treatment

$$\frac{\partial \rho}{\partial t} = -\nabla \cdot \rho \mathbf{u} \tag{1}$$

$$\frac{\partial \rho \mathbf{u}}{\partial t} = -\nabla \cdot (\rho \mathbf{u}\mathbf{u}) - \nabla p + \mathbf{j} \times \mathbf{B} \tag{2}$$

$$\mathbf{E} + \mathbf{u} \times \mathbf{B} = \eta \mathbf{j} \tag{3}$$

$$\frac{\partial p}{\partial t} = \nabla \cdot p\mathbf{u} - (\gamma - 1)\left(p \nabla \cdot \mathbf{u} + \eta \mathbf{j}^2\right) \tag{4}$$

$$\nabla \times \mathbf{B} = \mu_0 \mathbf{j} \tag{5}$$

$$\frac{\partial \mathbf{B}}{\partial t} = -\nabla \times \mathbf{E} \tag{6}$$

with the plasma density ρ, velocity $\mathbf{u}$, pressure p magnetic field $\mathbf{B}$, electric field $\mathbf{E}$, and resistivity η. This approximation does not consider kinetic effects, however, provides a relatively simple approximation which is valid on sufficiently large length scales.

2 The Bow Shock and the Magnetosheath

The magnetosphere is embedded in the solar wind, a high velocity stream of plasma (mostly protons and electrons) originating from the sun. Velocities of the solar wind are between 300 km/s and 1400 km/s with a typical value of about 500 km/s. Densities are between 1 and about 100 particles per cm^3 with a typical value of about $5\,cm^{-3}$ at 1 AU. The solar wind is relatively cold and the thermal and magnetic energy density are much smaller than the bulk kinetic energy. The magnetic field has a spiral shape and is usually closely aligned with the ecliptic plane.

However, the actual boundary conditions for the magnetosphere are not the unperturbed solar wind properties. The solar wind speed is much faster than the fast wave speed (fast mode) based on the MHD equations. This implies the formation of a shock in front of the Earth's magnetosphere which decelerates the solar wind to velocities slower than the fast mode speed. The principle is similar to the formation of acoustic shocks by a supersonic aircraft.

2.1 The Bow Shock

Let us first take a brief look at the bow shock properties because it is not directly the solar wind but the shocked (heated and compressed) solar wind plasma which forms the magnetospheric boundary conditions. The location of the dayside magnetospheric boundary is typically about 10 R_E (Earth radii). Thus the curvature and associated length scales are much larger than kinetic plasma scales such as the ion gyroradius for thermal protons. Therefore the up- and downstream conditions of such a shock are reasonably described by the MHD equations. The basic principle to model shocks and discontinuities in a fluid description uses the large scale conservation laws. To illustrate this let us assume a continuity equation of the form

$$\frac{\partial f}{\partial t} = -\nabla \cdot f\mathbf{u} \tag{7}$$

where f represents a quantity such as the mass or particle density and the equation implies that no mass is generated or annihilated. A stationary ($\partial/\partial t = 0$) solution across a one dimensional discontinuity implies that the particle flux $fu_n = const$ where u_n is the velocity normal to the one-dimensional boundary. In the case of a magnetized plasma one can write the ideal MHD equations ($\eta = 0$) in conservative form such as (7). These equations complemented with $\nabla \cdot \mathbf{B} = 0$ lead to a set of jump conditions (Rankine-Hugoniot conditions) for one-dimensional MHD discontinuities (e.g. Priest, 1987 [77]; Baumjohann and Treumann, 1997 [6]).

$$\mathbf{n} \cdot [\rho \mathbf{u}] = 0 \tag{8}$$

$$\mathbf{n} \cdot [\rho \mathbf{u}\mathbf{u}] + \mathbf{n} \left[p + \frac{\mathbf{B}^2}{2\mu_0} \right] - \frac{1}{\mu_0} \mathbf{n} \cdot [\mathbf{B}\mathbf{B}] = 0 \tag{9}$$

$$\mathbf{n} \cdot \left[\left(\frac{1}{2} u^2 + \frac{\gamma p}{(\gamma - 1)\, \rho} + \frac{1}{\mu_0 \rho} B^2 \right) \rho \mathbf{u} \right] - \frac{1}{\mu_0} \mathbf{n} \cdot [(\mathbf{u} \cdot \mathbf{B})\, \mathbf{B}] = 0 \tag{10}$$

$$\mathbf{n} \times [\mathbf{u} \times \mathbf{B}] = 0 \tag{11}$$

$$\mathbf{n} \cdot [\mathbf{B}] = 0 \tag{12}$$

Solutions of these conditions can be classified into discontinuities in which the velocity normal to the boundary u_n is constant $[u_n] = u_{nu} - u_{nd} = 0$ and shocks for which the normal velocity changes from the up- to the downstream region. Here the indices u and d indicate up- and downstream regions. The equations particularly imply that the normal magnetic field and the normal mass (or particle) flux are constant. They also imply the existence of a deHoffmann-Teller (dHT) frame, i.e., a frame of reference in which the tangential electric field $\mathbf{E}_t = -(\mathbf{u} \times \mathbf{B})_t$ is zero which implies that the plasma velocity is aligned with the magnetic field.

Introducing the angle θ between the incident magnetic field and the shock normal $\mathbf{n}$, the plasma compression $X = \rho_d/\rho_u$, and assuming the tangential velocity along the y direction such that $u_y/B_y = u_n/B_n$ the jump conditions can be combined to a single equation for the plasma compression X (e.g. Priest, 1987 [77])

$$\left(u_u^2 - X u_A^2\right)^2 \left[X c_s^2 + \frac{1}{2} u_u^2 \cos^2\theta \left\{ X(\gamma - 1) - (\gamma + 1) \right\} \right]$$
$$+ \frac{1}{2} u_A^2 u_u^2 X \sin^2\theta \left[(\gamma + X(2 - \gamma))\, u_u^2 - X u_A^2 \left((\gamma + 1) - X(\gamma - 1) \right) \right] = 0 \tag{13}$$

with $u_A = B_u/\sqrt{\mu_0 \rho_u}$ and γ is the ratio of specific heats. Using the jump conditions up- and downstream properties are related to the compression by

$$\frac{u_{nd}}{u_{nu}} = \frac{1}{X} \qquad \frac{p_d}{p_u} = X + \frac{\gamma - 1}{2} \frac{X u_u^2}{c_{su}^2} \left(1 - \frac{u_d^2}{u_u^2} \right)$$
$$\frac{u_{yd}}{u_{yu}} = \frac{u_u^2 - u_{Au}^2}{u_u^2 - X u_{Au}^2} \qquad \frac{B_{yd}}{B_{yu}} = \frac{u_u^2 - u_{Au}^2}{u_u^2 - X u_{Au}^2} X$$

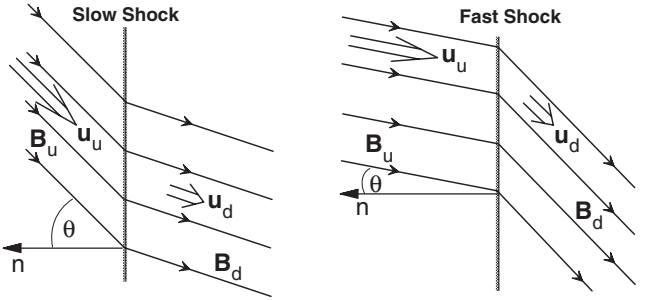


Fig. 6. Illustration of slow and fast shocks.

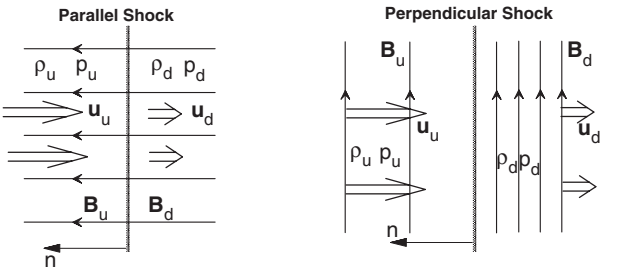


Fig. 7. Illustration of parallel and perpendicular shocks.

Equation (13) has 3 solutions. One of which has $X = 1$ and $u_u^2 = u_{Au}^2$ and represents a nonlinear Alfvén wave or so-called rotational discontinuity. The other two solutions are referred to as a slow and a fast shock (referring to the corresponding MHD wave solutions). Both solutions are compressive and imply $p_d > p_u$. The main difference appears in in terms of the tangential magnetic field and velocity. The fast shock solution implies $B_{yd} > B_{yu}$ and $u_{yd} > u_{yu}$ whereas the slow shock has $B_{yd} < B_{yu}$ and $u_{yd} < u_{yu}$ as illustrated in Fig. 6. For a fast shock the upstream velocity is faster than the upstream fast mode phase speed $c_f^2 = c_s^2 + u_{At}^2$ where $c_s^2 = \gamma k_B T/m_i$ is the sound speed and $u_{At}^2 = B_y^2/\mu_0\rho$ is the Alfvén speed based on the tangential magnetic field strength.

There are two special cases for the fast shock worth illustrating. The fast shock is called a parallel shock if the magnetic field is aligned with the shock normal and it is called a perpendicular shock if the field is perpendicular to the shock normal (Fig. 7). In the case of a parallel shock the downstream magnetic field is unaltered whereas it is strongly compressed in the case of a perpendicular shock.

The downstream properties are frequently given as a function of the upstream sonic Mach number $M = u_u/c_s$. Strong fast shocks which are typical for the bow shock have $M \gg 1$. In this case the downstream properties approach

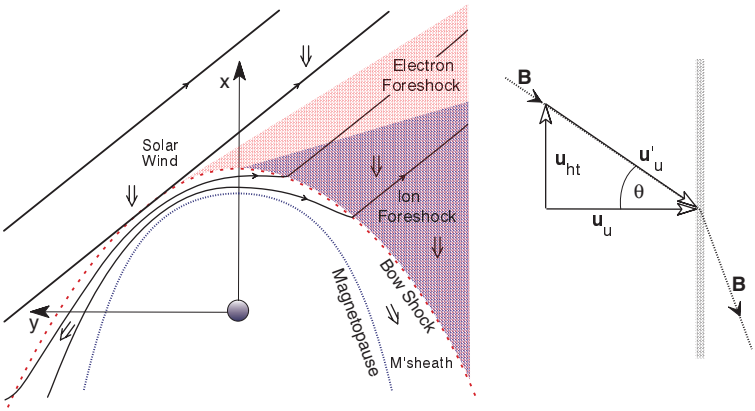


Fig. 8. Illustration of foreshocks and the deHoffmann-Teller frame.

$$\frac{\rho_d}{\rho_u} = 4 \quad \frac{u_{nd}}{u_{nu}} = \frac{1}{4} \qquad \frac{p_d}{p_u} = \frac{2\gamma M_u^2}{\gamma + 1}$$

The strongest magnetic field amplification occurs for a perpendicular shock with $B_d/B_u = 4$. For the bow shock one also use the terminology quasi-parallel or quasi-perpendicular shocks to indicate that a shock is almost parallel or almost perpendicular. Since the bow shock is curved there is almost always a location where the shock is almost perpendicular or almost parallel as illustrated in Fig. 8.

This discussion sets the framework for the bow shock. The distance of the bow shock from the Earth is approximated by

$$R_{bs} = \left(1 + 1.1\frac{n_{sw}}{n_{msh}}\right) R_{mp}$$

Since the bow shock is curved the magnitude of the normal velocity u_{nu} changes according to $u_{nu} = \mathbf{n}_{bs} \cdot \mathbf{u}_{sw} = u_{sw} \cos\varphi$ where φ is the angle between the shock normal and the Sun-Earth line. This also implies that the bow shock is limited in size because the normal velocity becomes too small to generate a shock when φ is sufficiently large.

To understand the detailed structure and dissipation in the bow shock a kinetic treatment is necessary (e.g. Stone and Tsurutani, 1985 [107]; Burgess, 1995 [20]). Such studies are performed usually with hybrid (kinetic ions and fluid electrons) simulations or full particle simulations. The goals of these studies are to understand the dissipation mechanisms, the associated shock structure, and the acceleration of particles in the shock.

The fastest speed with which information can travel in MHD is the fast mode speed which implies that no information can travel upstream of the bow shock withing the MHD framework. However, in a kinetic treatment of shocks particles can in fact travel upstream of the bow shock. Thus energetic

solar wind particles which are scattered in the bow shock or particles which are accelerated in the bow shock can travel into the upstream region. The particles form the electron and ion foreshocks (note that ions are slower for the same energies than electrons and such that the ion foreshock is Earthward of the electron foreshock). These regions are highly turbulent because of the particle beams and associated plasma instabilities, and this turbulence is subsequently carried back across the bow shock.

2.2 Magnetosheath Flow and Structure

The magnetosheath is the region between the bow shock and the magnetosphere and represents the shocked solar wind plasma. The plasma flow in the magnetosheath is mostly tangential to the magnetosphere. The magnetosheath is fairly turbulent (e.g. Engebretson et al., 1991 [31]). Several important source for this turbulence are

- Large scale magnetic field fluctuations associated with quasi-parallel shocks (Fig. 9): Small changes in the upstream magnetic field orientation generate large changes in the magnetosheath field because of the amplification of the tangential magnetic field component (Lin et al., 1996 [49]).
- Kinetic processes at quasi-parallel shocks (e.g. Krauss-Varban and Omidi, 1993 [41]; Scholer et al., 1993 [94].
- Since the conditions downstream of the bow shock are different any MHD wave or other solar wind perturbation (such as a density variations) launches various MHD waves at the bow shock which travel into the magnetosheath (e.g. Borovsky and Funsten, 2003 [17]).
- Fast mode waves launched from the bow shock travel through the magnetosheath and are partially transmitted and mostly reflected at the magnetopause (e.g. Otto, 1995a [63]). The reflected waves can bounce between the bow shock and the magnetopause.
- Driven by electric currents, particle beams, or pressure anisotropy various plasma instabilities operate in the magnetosheath close to the bow shock. The most prominent of these are whistler, lower hybrid, ion-acoustic, ion-cyclotron, and mirror modes (e.g. McKean et al., 1992 [51]; Treumann and Baumjohann, 1997 [109]).

Another typical property of the magnetosheath is the so-called plasma depletion layer in the subsolar magnetosheath particularly for northward IMF orientation (Midgley and Davis, 1963 [53]; Zwan and Wolf, 1976 [117]; Anderson et al., 1997 [3]). The cause for this plasma depletion is the following. Let us assume for a moment that the magnetosphere is a two-dimensional perfectly conducting obstacle of cylindrical shape. In this case plasma can still flow around the obstacle in a 2D stagnation flow, however, the magnetic flux cannot cross the obstacle and is piled up in front of the obstacle (Fig. 10).

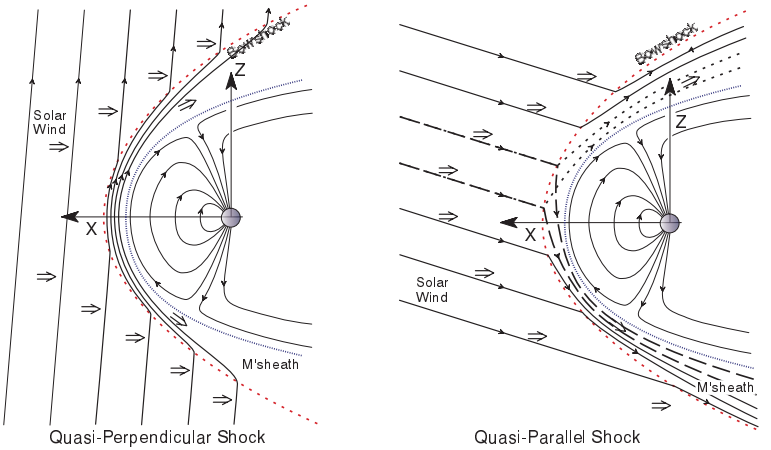


Fig. 9. Illustration of quasi-perpendicular and quasi-parallel bow shock situations.

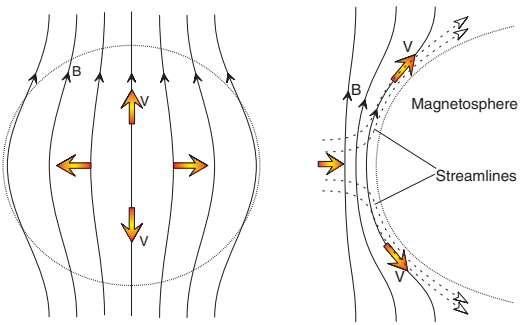


Fig. 10. Left: 3D stagnation flow in a view from the sun onto the dayside magnetopause (circle). Right: 2D flow and flux pile up.

The real magnetosphere is three-dimensional and the stagnation flow has two degrees of freedom (i.e. east-west and north-south). However, the magnetic flux can be transported around the magnetosphere only by plasma flow perpendicular to the direction of the magnetic flux, e.g., the east-west components of the stagnation flow when the magnetic field is along the north-south direction. This generates a pile-up of the magnetic flux and because of the increased magnetic pressure the corresponding plasma pressure and density is reduced causing the plasma depletion region. Note that for southward IMF magnetic reconnection allows a different mode for the flux transport as will be discussed in the next section.

3 The Magnetopause

The magnetopause is usually defined as the boundary between the magnetosphere and the magnetosheath (the shocked solar wind). It should be noted that this definition has a problem in regions where the solar wind plasma has access to the magnetosphere. Thus for practical purposes it may sometimes be more convenient to define the magnetopause as the region of highest current density between the magnetosphere and the magnetosheath with a mixture of magnetospheric and magnetosheath particles.

An important aspect with respect to the magnetospheric boundary is the magnetic topology (connection). There are three different types of magnetic connections for magnetic field lines close to the magnetopause (Fig. 11):

- Closed geomagnetic field lines have both 'end' points in the Earth
- Open magnetic field lines have one foot point on the Earth and connect with the other side to the solar wind
- IMF field lines are not connected to the Earth

The magnetopause controls the transport of mass, momentum, energy, and magnetic flux into the magnetosphere. This transport and the circulation of magnetic flux is closely related to the magnetic topology and the change of this topology, i.e., the generation of newly opened magnetic flux. Open magnetic flux is generated by magnetic reconnection and which will be discussed later in this section. First we will focus on basic properties and observations of magnetopause physics.

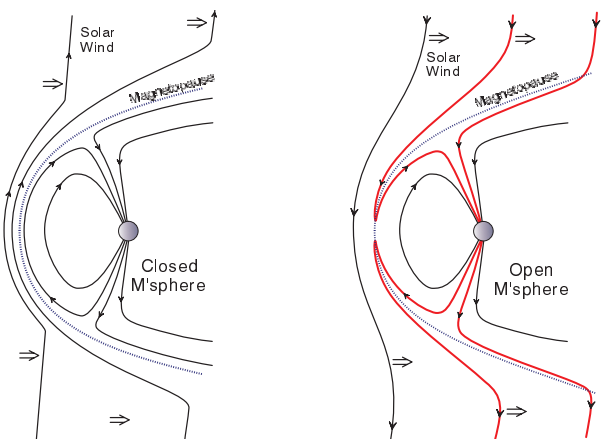


Fig. 11. Sketch of closed and open magnetospheric configurations

3.1 Basic Properties and Observations

Magnetopause Shape

The shape of the magnetopause boundary is controlled by the solar wind dynamic pressure and the internal magnetospheric pressure which is mostly the magnetic pressure (e.g. Walker and Russel, 1995 [115]). The solar wind dynamic pressure is $p_{dyn} = \kappa m_p n_{sw} \left(\mathbf{u}_{sw} \cdot \mathbf{n}_{mp}\right)^2$ where m_p = proton mass, n_{sw} = solar wind particle density, $\mathbf{u}_{sw}$ =solar wind velocity, $\mathbf{n}_{mp}$ = magnetopause normal, and κ is a coefficient of order unity which takes into account that it is not directly the solar wind but the shocked solar wind pressure which is exerted onto the magnetosphere. The pressure on the magnetospheric side is mostly the magnetic field pressure $p_{msp} = B^2_{msp}/2\mu_0$. Defining the flaring angle θ to be the angle between the the solar wind speed and the direction tangential to the magnetopause pressure balance implies $\kappa m_p n_{sw} u^2_{sw} \sin^2\theta = B^2_{msp}/2\mu_0$. Flaring is the increase of the magnetopause distance from the sun-Earth line for increasing distance from the subsolar point (the point where the magnetopause is closest to the sun) as illustrated in Fig. 12.

Close to the subsolar point the geomagnetic field can be approximated by the dipole field which yields

$$m_p n_{sw} u^2_{sw} = \frac{K B^2_E}{2\mu_0 R^6_{mp}}$$

where B_E =dipole field at the Earth's equator and K is a constant that includes the effects covered by κ as well as any errors due to the dipole field

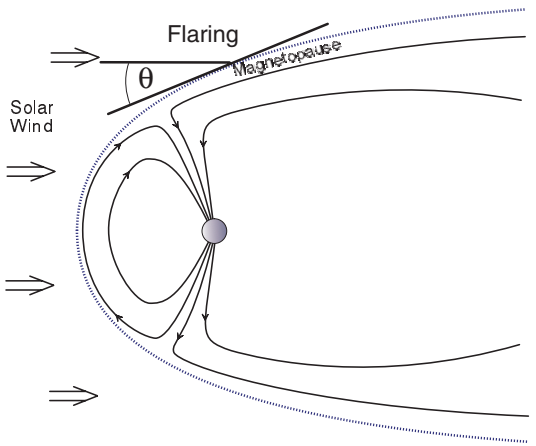


Fig. 12. Illustration of magnetospheric flaring.

approximation. Thus the stand-off distance of the magnetopause is

$$R_{mp} = \left(\frac{K B_E^2}{2\mu_0 m_p n_{sw} u_{sw}^2} \right)^{1/6}$$

For typical solar wind parameters $n = 5\,\mathrm{cm}^{-3}$, $u_{sw} = 400\,\mathrm{km}/s$, $B_E = 3 \cdot 10^4\,\mathrm{nT}$, and $K = 2$ the stand-off distance is $R_{mp} \approx 10\,\mathrm{R_E}$. For extreme solar wind conditions ($n = 100\,\mathrm{cm}^{-3}$, $u_{sw} = 1600\,\mathrm{km}/s$) the magnetopause stand-off distance may actually be around $R_{mp} = 5\,\mathrm{R_E}$.

Flaring implies the widening of the magnetospheric cavity with increasing distance from the subsolar point. The flaring angle is

$$\sin^2\theta = \frac{B_{msp}^2}{2\mu_0 \kappa m_p n_{sw} u_{sw}^2}$$

implying a decreasing flaring angle with a decreasing internal magnetic pressure. At large distances on the tailward side the thermal solar wind pressure eventually becomes comparable or larger than the normal dynamic pressure. Flaring seizes when the magnetic pressure on the magnetospheric side approximately equals the thermal solar wind pressure.

Magnetopause Structure

The local magnetopause structure is often characterized as a tangential or a rotational discontinuity (Fig. 13), e.g. Sonnerup et al. (1981) [105], Paschmann et al. (1986) [75], and Baumjohann and Treumann (1997) [6]. The terminology refers to MHD discontinuities where the normal magnetic field $B_n = 0$ for the tangential discontinuity and $B_n = const \neq 0$ for a rotational discontinuity. In other words the tangential discontinuity implies a

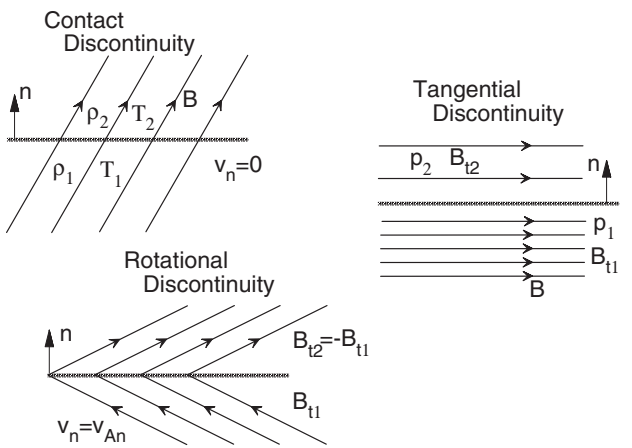


Fig. 13. MHD discontinuities.

closed magnetic boundary, there is no plasma flow across the boundary, and the total pressure is balanced $\left[p + B_t^2/\left(2\mu_0\right)\right] = 0$.

For the rotational discontinuity the normal magnetic field is nonzero such that the geomagnetic field is connected to the IMF. In the deHoffmann-Teller frame ($\mathbf{E} = -\mathbf{v} \times \mathbf{B} = 0$ or $\mathbf{v} \| \mathbf{B}$) the velocity normal to the discontinuity satisfies $u_n^2 = u_{An}^2$ with $u_{An} = B_n/\sqrt{\mu_0 \rho}$. Other important relations for the rotational discontinuity are

$$B_{td} = -B_{tu} \qquad p_d = p_u \text{ and } [u_t] = \pm [v_{At}]$$

where the last relation relates the difference in the up- and down-stream velocities to the difference in the up- and downstream Alfvén speeds tangential to the discontinuity. This relation is also know as the Walen relation and frequently used to identify Alfvén waves and in particular rotational discontinuities in satellite data. Note that the presence of a rotational discontinuity at the magnetopause is often interpreted as an expression of magnetic reconnection because reconnection of IMF and geomagnetic field generates an open magnetopause.

Except for these rather coarse models in terms of fluid discontinuities there are very few models of the actual structure of the magnetopause boundary. Early attempts to to explain the magnetopause structure and width in terms of the difference of electron and ion gyro radii were not self-consistent and failed to explain the thickness of the magnetopause which is typically larger than the ion gyro radius of typical particles. The reason for the absence of analytic models of the magnetopause is likely in the dynamic nature of this boundary. Various processes such as magnetic reconnection, shear flow instabilities, and continuous nonlinear perturbations from the magnetosheath side continuously alter the boundary and make an equilibrium or steady state solution difficult.

Magnetopause Motion, Thickness, and Plasma Properties

Observations show that the magnetopause is always in motion with typical in- and outward velocities of several 10 km/s (e.g. Russel, 1990 [82]; Russel, 1995b [84]). Therefore a spacecraft traverses the MP in general not because of the SC velocity but because of this rapid magnetopause motion. The causes for this motion are various but most prominently variations in the upstream solar wind (the dynamic pressure) and changes in the magnetic field orientation which can lead to increasing or decreasing reconnection rates.

The magnetopause thickness is typically 800 km but can vary between about 100 km and 2000 km. In satellite data the magnetopause structure is most easily identified close to the subsolar region and more turbulent and less obvious at high latitudes and the flanks of the magnetosphere where the velocity on the magnetosheath side is large.

The typical magnetopause boundary layer structure for inward SC trajectory is the following (Fig. 14): Plasma and magnetic field in the mag-

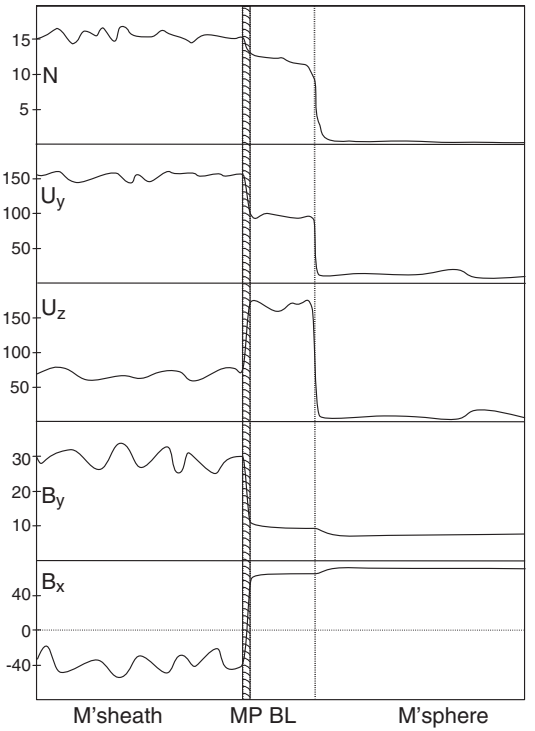


Fig. 14. Illustration of a typical magnetopause observation

netosheath are fairly disturbed. Then the magnetopause is identified as a sudden large change of the magnetic field orientation (and magnitude) in the direction tangential to the boundary. On the Earthward side of the boundary the magnetic field turns predominantly into the northward z direction. At the magnetopause the density may decrease slightly and the plasma consists of a mixture of magnetospheric (energetic) and magnetosheath (cold) particles. At the same time the plasma velocity changes strongly (rotation and acceleration). The layer of accelerated plasma flow and intermediate densities is called the low-latitude boundary layer (LLBL). The width of the LLBL is typical several thousand km. It is strongly varying and at times it may be entirely absent. The average width of the LLBL increases for increasing distance from the subsolar region.

One of the problems with satellite observations is that the observations represent a time series of data along a one-dimensional trajectory in a usually fairly complex system. To identify the orientation of the magnetopause boundary a method called variance analysis is applied to the satellite data (Sonnerup and Cahill, 1967 [102]; Sonnerup et al., 1987 [103]; Paschmann and Daly, 1998 [73]). This analysis assumes that the boundary is relatively thin

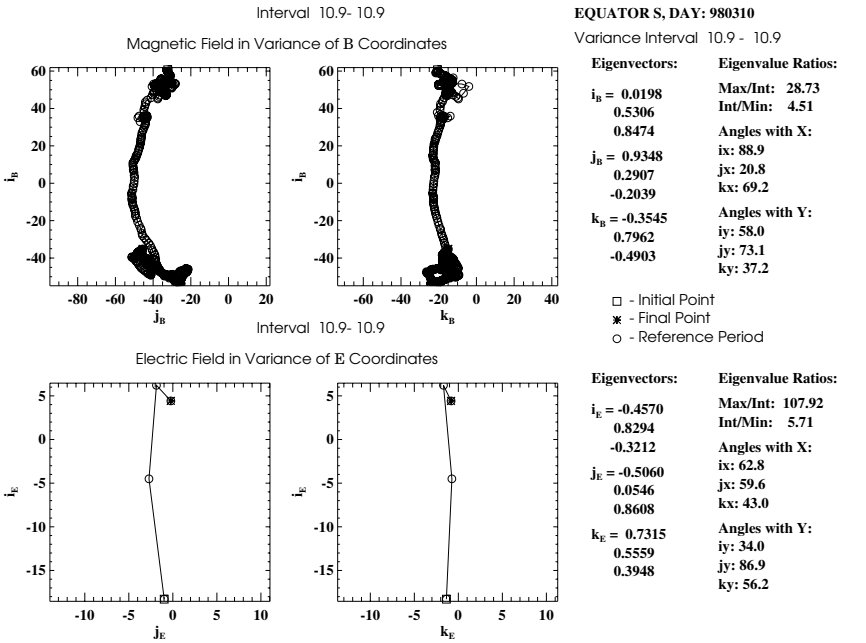


Fig. 15. Hodogram of the magnetic field and the electric field in variance of B and E coordinates for data from the Equator S satellite.

is approximately one-dimensional (that is plasma properties and magnetic field vary only in the direction normal to the boundary. For a thin boundary $\nabla \cdot \mathbf{B} = 0$ implies $B_n = const$. Thus the direction for which the magnetic field has the least variation is the direction normal to the boundary. This is formalized in a variance analysis where a series of magnetic measurements $B_\mu^{(i)}$ (μ is the Cartesian component and i indicates the time series index) is used to generate a variance matrix

$$M_{\mu\nu}^B = \frac{1}{N}\sum_{i=1}^{N} B_\mu^{(i)} B_\nu^{(i)} - \left[\frac{1}{N}\sum_{i=1}^{N} B_\mu^{(i)}\right]\left[\frac{1}{N}\sum_{i=1}^{N} B_\nu^{(i)}\right].$$

The eigenvector of this matrix with the smallest (in magnitude) eigenvalue identifies the direction with the smallest magnetic field variation, i.e., the normal direction. The same principle can be applied to the electric field data (Sonnerup et al., 1987 [103]) since the largest changes in velocity and magnetic field occur for the components tangential to the magnetopause the maximum variance of the electric field can be considered to indicate the normal direction.

Figure 15 shows the results of variance analysis for Equator S data of a magnetopause crossing on the dawnside flank of the magnetosphere. Here **i**

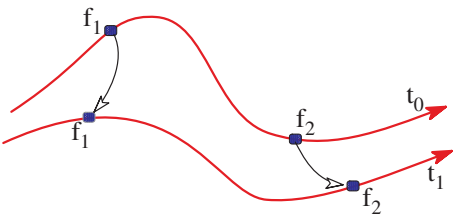


Fig. 16. Illustration of line conservation

refers to the maximum variance direction and **k** to the minimum variance direction. It is seen that the minimum variance eigenvector direction of the magnetic field is closely aligned with the maximum variance eigenvector of the electric field.

3.2 Processes at the Magnetopause

A central problem of magnetospheric dynamics and space weather is the transport of mass momentum, and energy through the magnetopause boundary into the magnetosphere. For individual particles it is straightforward that they carry out a gyro motion around a magnetic field line combined with with various drifts caused by gradients and curvature in the magnetic field. However, these drifts are tangential to the magnetopause and do not provide any significant particle transport across the boundary. More formally ideal Ohm's law $\mathbf{E}+\mathbf{u}\times\mathbf{B}=0$ implies the so-called frozen-in condition which states that the magnetic flux through any closed contour $\Phi_C = \int_C \mathbf{B}\cdot d\mathbf{s}$ moving with the plasma velocity **u** is constant in time. Equivalently, any two fluid elements f_1, f_2 are always connected by the same magnetic field line (see Fig. 16) if they were connected at one time by this field line (defined by the direction of the magnetic field at any moment in time). In other words a magnetic field line can be identified by the motion of such fluid elements. This can be extended to a more general form of Ohm's law when the bulk velocity **u** is replaced by the electron velocity $\mathbf{u}_e$.

Thus plasma on interplanetary field lines (without connection to the Earth) cannot penetrate into the magnetosphere without violating ideal MHD or Hall MHD (for a more general form of Ohm's law). Since the magnetospheric plasma is collisionless there is no large scale resistivity such that the violation of Ohm's law can only occur localized on very small scales. Note that momentum and energy can still be transferred into the magnetosphere through waves or viscous coupling. The most important processes for mass, momentum, or energy transfer into the magnetosphere are thought to be

- Magnetic reconnection
- Viscous interaction
- Pressure pulses and impulsive penetration

In the following we will discuss examples and properties of these processes.

Magnetic Reconnection

During periods of southward IMF magnetic reconnection connects closed geomagnetic field with interplanetary magnetic field generating newly opened magnetic flux (Fig. 11). Along the newly opened field particles can freely enter and leave the magnetosphere. The newly opened field is swept with the solar wind along the magnetosphere and magnetic flux accumulates at the tail boundary. The accumulation of magnetic flux in the magnetotail lopes increases the size of the tail magnetosphere and magnetic energy is stored in the lobes of the magnetotail. Thus reconnection at the dayside magnetopause is important for the energy budget of the entire magnetosphere. The subsequent processes in the magnetotail will be addresses in the section on the magnetotail.

Magnetic reconnection implies the new connection of magnetic flux and therefore requires a violation of Ohm's law albeit in a very small region. In terms of analytic theory there are two main analytic approaches to the process of magnetic reconnection. One uses a steady state assumption for the local reconnection geometry and the other one treats reconnection as an instability (tearing mode) and addresses questions like the onset of reconnection. This instability aspect is more relevant for magnetotail dynamics.

Reconnection Models: There are currently no self-consistent general models of magnetic reconnection. Typically reconnection models assume a steady state which implies $\partial B/\partial t = 0$ or $\nabla \times \mathbf{E} = 0$. In two dimensions with $\partial/\partial z = 0$ this implies $E_z = const$, e.g. Vasyliunas (1975) [114] and Priest and Forbes (2000) [78]. The basic geometry of the earliest reconnection models (Parker, 1957 [70]; Sweet, 1958 [108]; Petschek (1964) [76]) is illustrated in Fig. 17. Here the region indicated in red is called the diffusion region. In simplified sketches of reconnection this region is contracted to a single point (as for instance in Fig. 11) and is addressed as the x point or x line (assuming invariance along the z direction) because of the x type magnetic field configuration at this point. The magnetic field lines which connect to the diffusion

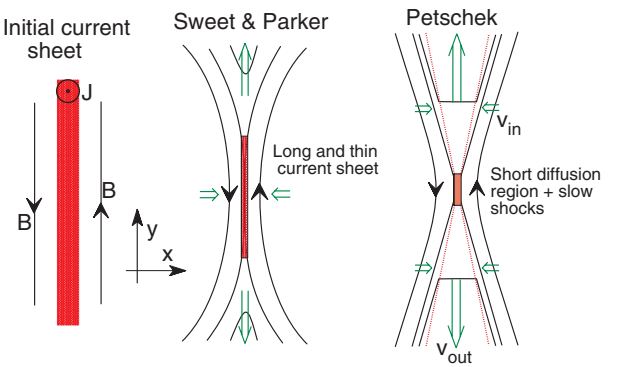


Fig. 17. Initial current sheet and basic reconnection geometries.

region are called separatrices (in three dimensions these are separatrix surfaces). The separatrices approximately separate the in- and outflow regions. The inflow regions are to the left and right (Fig. 17) of the diffusion region and plasma converges toward the central region. Plasma is jetting away from the diffusion region in the outflow regions.

The z component of the electric field is the same in the inflow, outflow, and diffusion regions. Using the simple form of resistive Ohm's law (3) the electric field in the in- and outflow regions is given by $-(\mathbf{v}\times\mathbf{B})_z$ while it is approximately ηj_z in the diffusion region. Note that the electric field represents also the rate for the magnetic flux transport into and out of the diffusion region and is therefore the magnetic reconnection rate.

It is useful to measure the reconnection rate in units of a typical electric field in order to re-scale reconnection for various plasma configurations and parameters. Considering a typical magnetic field strength of B_0 (the asymptotic magnetic field in the inflow regions), a typical density ρ_0 which yields an Alfv́en speed of $u_A = B_0/\sqrt{\mu_0\rho_0}$, the reference electric field magnitude is $E_0 = v_A B_0$. Measuring the reconnection rate in units of this typical electric field yields a dimensionless reconnection rate $\epsilon_r = E_z/E_0$. Since the electric field in the inflow region is $E_z = v_{in}B_0$ the dimensionless reconnection rate can also be expressed as $\epsilon_r = v_{in}/v_A$.

In steady reconnection models it is usually assumed that the density is approximately constant in the diffusion region. In this case inflow and outflow velocities satisfy $v_{in}/v_{out} = d/L$. where d is the width and L is the length of the diffusion region. Finally simple steady state models imply that the outflow velocity is approximately $v_{out} \approx v_A$ yielding $B_{out} \approx \epsilon_r B_0$.

In the earliest models (Parker, 1957 [70]; Sweet, 1958 [108]) resistive diffusion of the magnetic field in a mostly one-dimensional current sheet was assumed and reconnection scaled as $\epsilon_r = S^{-1/2}$ with the magnetic Reynolds number $S = v_A L/\eta$. Many space plasma systems are either collisionless with $\eta \approx 0$ or have have a very large magnetic Reynolds number such that the reconnection rate is extremely small. Petschek suggested a model which based on the evolution of slow shocks which bound the outflow region. In this case the reconnection rate scales with $\sim \text{const}/\ln S$ yielding more realistic values of 0.01 to 0.1 even for very large magnetic Reynolds numbers.

Subsequently many other reconnection models have been suggested to account for different configurations and to fix shortcomings of these early models (e.g. Otto, 1995b [64]; Schindler, 1995 [88]; Priest and Forbes, 2000 [78]). For instance the typical magnetopause structure implies much higher densities and lower magnetic field strength on the magnetosheath side. Also there is significant plasma flow on the magnetosheath side and the magnetic field is never really antiparallel on the two sides of the magnetospheric boundary.

During recent years many of these aspects have been studied also by numerical simulation (e.g. Lee, 1995 [44]; La-Belle-Hamer et al., 1995 [43]; Otto et al., 1995 [68]; Otto, 1995a [63]; Otto, 1989 [65]). Numerical models show

that constant small resistivity tends to yield reconnection with long thin current sheets similar to the Sweet-Parker model while current dependent or otherwise localized resistivity yields short diffusion regions and fast reconnection. Particularly noteworthy is a series of numerical models which include the Hall term in Ohm's law (Birn et al., 2001 [9]; Shay et al., 2001 [98]; Otto, 2001 [66]). Independent of the specific physics in different simulations (Hall MHD, Hybrid with kinetic ions and fluid electrons, fully electromagnetic) the results show approximately the same reconnection rate and plasma configuration. Thus the inclusion of the Hall term yields Petschek-like fast reconnection for certain simple initial configurations.

Observation of Magnetopause Reconnection: The observational verification of reconnection is difficult because a satellite provides only point measurements of usually complex plasma structure. Two major signatures are used to identify reconnection at the magnetopause.

(a) On macro scales reconnection should generate open magnetic field which implies that the magnetopause is at time approximately a rotational discontinuity. In this case measurements of the plasma velocity and magnetic field should satisfy approximately the so-called Walen relation

$$\Delta\mathbf{u} = \pm\Delta\mathbf{v}_A = \pm\Delta\frac{\mathbf{B}}{\sqrt{\mu_0\rho}}.$$

where $\Delta\mathbf{u} = \mathbf{u} - \mathbf{u}_{ref}$ with a measured velocity $\mathbf{u}$ and a reference measurement $\mathbf{u}_{ref}$. While early tests provided only few events which satisfied this relation later measurements with better temporal resolution provided many cases of thin magnetopause current layers which approximately satisfy the Walen relation (Sonnerup et al., 1981 [105]; Paschmann et al., 1986 [75]; Gosling et al., 1990 [35]). An additional test for the presence of a stationary magnetopause structure is the presence of a dHT frame (Sonnerup et al., 1990 [104]). Since the electric field is assumed constant there should be a reference frame in which the electric field is almost zero. An example which shows an excellent dHT frame but a poor Walen test is shown in Fig. 18. The events typically are of short duration (few 10 seconds) indicating thin layers, show a mixture of magnetosheath and magnetospheric plasma, and a occur mostly for strong southward IMF.

(b) The second class of events are so-called magnetic flux transfer events (FTE's) (Haerendel et al., 1978 [36]; Russel and Elphic, 1978 [81]; Paschmann et al., 1982 [74]; Elphic, 1990 [29]; Elphic, 1995 [30]). They show typically a strong bipolar variation of the magnetic field component normal to the magnetopause $\mathbf{B}_n$ together with a mixture of magnetosheath and magnetospheric electron populations. Various other typical properties are a correlation with southward IMF, an increase of total pressure and total magnetic field, a strong rotation of the field in the core of the event, and a good dHT frame. The typical duration is one to few minutes with a repetition rate of about 8 minutes. An example of FTE data is shown in Fig. 19. The distribution of FTE's indicates the subsolar region as their source region. Furthermore

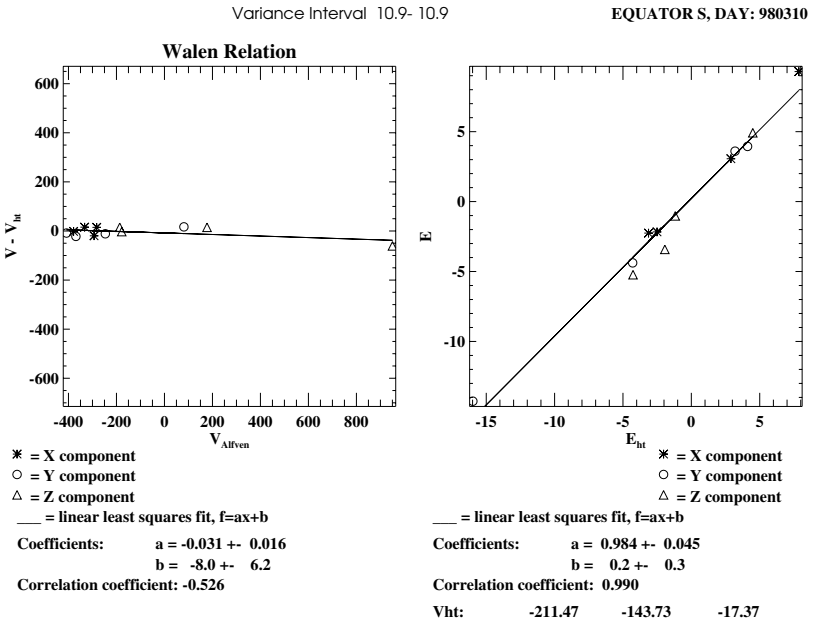


Fig. 18. Tests of the dHT frame and the Walen relation for the data set shown before.

typical amplitudes and scale sizes of FTE's require fast reconnection, i.e., with the Petschek reconnection rate.

Magnetopause Reconnection Models: The first explanation of FTE's was suggested by Russel and Elphic (1978) [81] who assumed that patchy magnetic reconnection generates a magnetic flux tube or rope which connects the magnetospheric and magnetosheath sides and thus contains magnetosheath and magnetospheric particles. Magnetic field draped around this tube can generate the bipolar B_n signature as the flux tube moves along the magnetopause boundary as illustrated in Fig. 20.

There are various alternative models to explain FTE signature. The most basic possibility is a short duration reconnection pulse in a two-dimensional approximation (Scholer, 1988 [93]; Southwood et al., 1988 [106]). This process is illustrated as a result of a 2D MHD simulation in Fig. 21. Here reconnection generates a plasma bulge which and the magnetic field around this bulge will generate a bipolar B_n signature.

Another model (Lee and Fu, 1985 [45]; Lee and Fu, 1986 [46]; Lee, 1995 [44]) employs reconnection at multiple x lines and the magnetic islands moving along the magnetopause will also cause a bipolar B_n signature. Both single x line reconnection model and multiple x line reconnection have also been simulated in three-dimensional models, i.e., with a finite (small) length of the reconnection region (Fu et al., 1990 [34]; Otto, 1990 [61]; Schindler

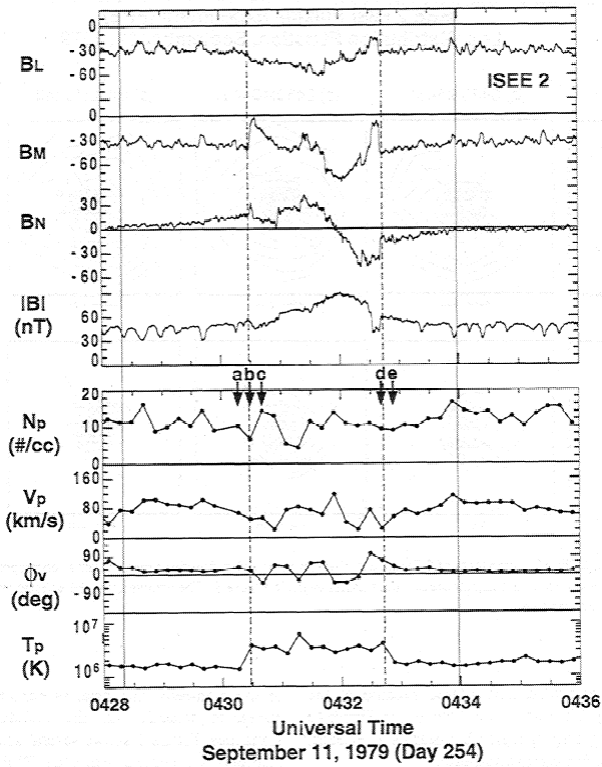


Fig. 19. Example of FTE data after Russel and Elphic (1978) [81].

and Otto, 1990 [92]; Ma et al., 1984 [50]). While the qualitative signatures are rather similar it appears that multiple x line reconnection and cases with relatively short x lines (3D) generate stronger and more realistic signatures. Currently it is unresolved whether reconnection occurs pulsed along a rather long x line along much of the dayside magnetopause or at multiple patches (Nishida, 1989 [58]; Otto, 1991 [62]; Otto, 1995a [63]) distributed of the subsolar region of the magnetopause as illustrated in Fig. 22.

Starting from very simple one-dimensional initial conditions, three-dimensional MHD simulations (Otto, 1989 [65]) demonstrate that reconnection initiated at a single reconnection site does not remains localized. Rather the initial single reconnection site decays into multiple reconnection sites which move along the current sheet.

Figure 23 shows the parallel electric field and the total pressure in the y, z plane at an early and a late time from a three-dimensional MHD simulation. Both quantities are integrated perpendicular to the current sheet (along x) to provide the global information which cannot be captures by a single cut through the 3D system. The parallel electric field is chosen because it

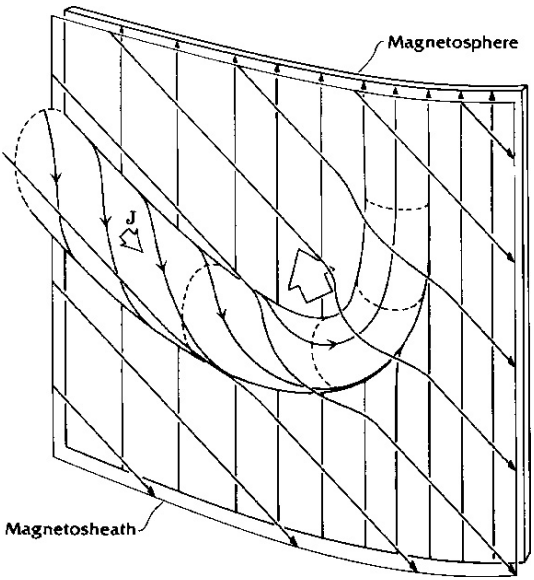


Fig. 20. Sketch of a magnetic flux rope and the magnetic field draping from Russel and Elphic (1978) [81].

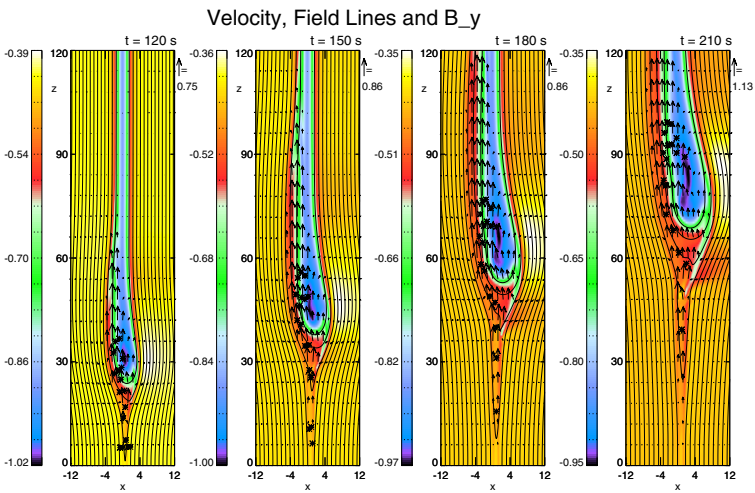


Fig. 21. 2D simulation of localized reconnection.

represents the individual diffusion regions (reconnection sites) and maxima in the total pressure represent the locations with FTE-like properties. The initial configuration is a one-dimensional current sheet and reconnection is triggered only at single location. The figure demonstrates that reconnection is not confined to the initial onset location and FTE signatures develop in

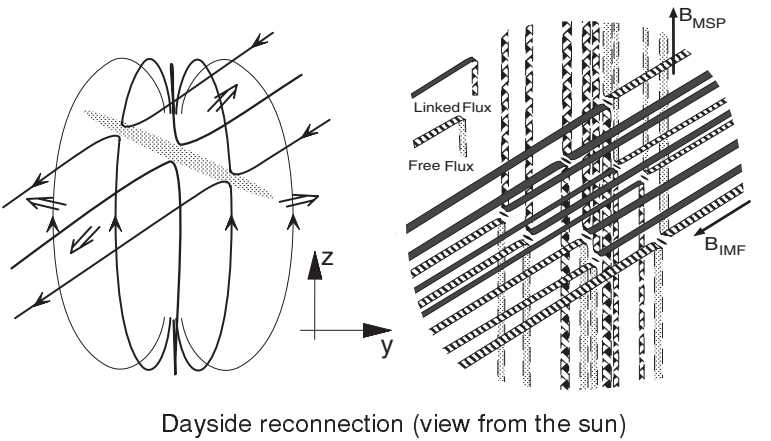


Fig. 22. Sketch of reconnection in a single reconnection region (left) or at multiple small reconnection patches (right) in a view from the sun onto the dayside magnetopause.

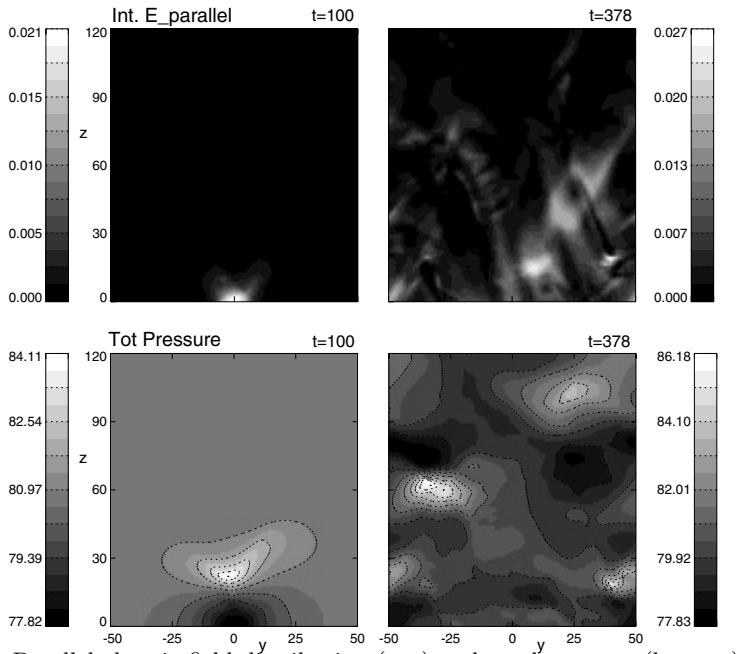


Fig. 23. Parallel electric field distribution (top) and total pressure (bottom) from 3D MHD simulations of magnetic reconnection at an early (left) and a late (right) time of the simulations.

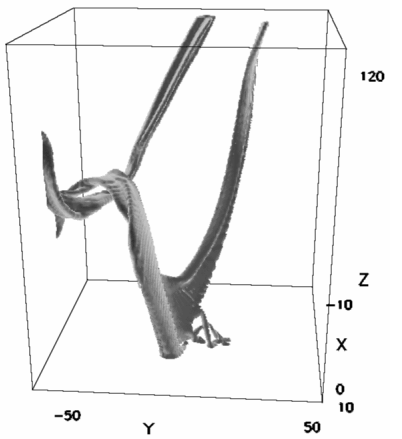


Fig. 24. Example of magnetic flux ropes from a 3D MHD simulation.

much of the simulation domain. The magnetic field configuration in Fig. 24 shows two flux tubes elbow-like interlinked in the region of the total pressure maximum (time $t = 378$).

Viscous Interaction

From the early days of magnetospheric physics it has been presumed that viscosity at the magnetospheric boundary may drive convection in the magnetosphere (Axford and Hines, 1961 [5]). This is in particular evident from ionospheric convection. There is tailward convection in the polar cap at all times and only for strongly northward IMF two additional convection cells develop which show some sunward flow driven by high latitude (cusp) reconnection. This indicates that a portion of the magnetospheric convection is always driven by the magnetosheath flow and it is mostly accepted that about 10 to 20 kV of the polar cap potential (Fig. 25) may be attributed to viscous processes at the magnetospheric boundary. In other words some of the tailward flow in particular close to the center of the convection cells may be on closed geomagnetic field lines.

In the outer magnetosphere the presence of the LLBL (Fig. 26) is indicative for viscosity driven convection. The LLBL consists of a mixture of magnetospheric and magnetosheath particles and plasma is flowing tailward although not quite as fast as in the adjacent magnetosheath. It is not yet clear whether the LLBL is entirely on closed field lines or is actually a mixture of closed and open magnetic field (Newell and Meng, 2003 [57]). On the dayside boundary magnetic reconnection may generate this boundary layer. The average width of the LLBL increases away from the subsolar region to about 0.5 R_E close to the terminator. A diffusion coefficient of $D = 10^9\,\mathrm{m^2 s^{-1}}$ is

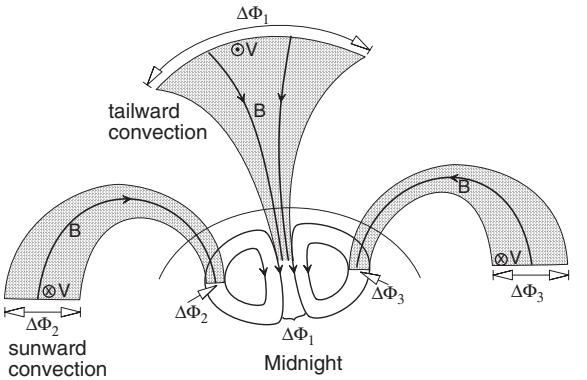


Fig. 25. Sketch of ionospheric convection and cross polar cap potential.

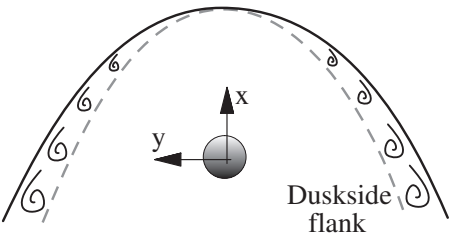


Fig. 26. Sketch of the LLBL in the equatorial plane of the magnetosphere.

required (Sonnerup, 1980 [101]) to account for the LLBL quantitatively by mass and momentum diffusion.

The presence of viscosity in a fluid implies that momentum can be transported in a direction transverse to the actual flow. However, as pointed out the magnetospheric plasma is highly collisionless meaning there are no classical collisions of particles in most of the magnetosphere. This leaves two main physical mechanisms which may account for the viscous coupling.

Microinstabilities: The magnetosphere has rather thin boundaries. These boundaries imply large gradients in many plasma properties such as density, temperature, or magnetic field. The free energy either directly due to the strong gradients or caused by electron/ion beams at the boundary can cause various instabilities. The effect of such instabilities is to relax the configuration which caused the instability which means to reduce the gradients or the fast relative motion of electrons and ions. Therefore they will cause diffusion of mass and momentum or friction similar to actual collisions of particles. Various microinstabilities such as lower hybrid drift modes, ion acoustic modes, ion cyclotron, etc. have been suggested to account for viscous interaction (La Belle and Treumann, 1988 [42]; Treumann and Baumjohann, 1997 [109]). However, while there are models to evaluate the resulting trans-

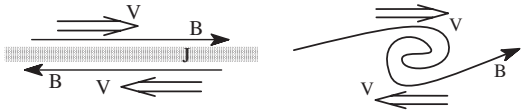


Fig. 27. Sketch of KH evolution in the presence of a magnetic field.

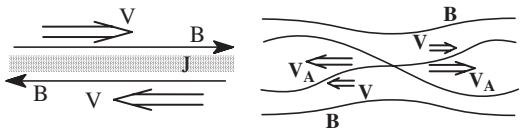


Fig. 28. Sketch of magnetic reconnection in the presence of sheared plasma flow.

port or diffusion coefficient, there are no self-consistent models of the full nonlinear coupling. This is an important point because of the rather large width of the LLBL close to the terminator.

Kelvin Helmholtz (KH) Waves: An alternative mechanism for the formation of the LLBL is the Kelvin Helmholtz instability (Chandrasekhar, 1961 [21]). This instability is present in many situations where two fluids stream relative to each other. Here the magnetosheath plasma (solar wind flow) moves fast relative to the plasma on the magnetospheric side which is almost at rest. Numerical simulations have demonstrated that the KH can in principle operate at the magnetospheric flanks (Miura and Pritchett, 1982 [56]; Miura, 1982 [54]) and it is able to transport energy and momentum from the magnetosheath into the magnetosphere (Miura, 1984 [55]). However, there are two aspects worth considering.

The KH mode is stabilized by a variety of physical effects such as viscosity, surface tension, or a magnetic field aligned with the plasma flow. In the latter case the magnetic field is deformed by the plasma flow (Fig. 27). This deformation requires additional energy such that the KH mode is stabilized (Chandrasekhar, 1961 [21]; Chen et al., 1997 [24]; Otto and Fairfield, 2000 [67]). In fact if the magnetic field energy density is higher than the energy in the shear flow the KH mode cannot operate at all.

Vice versa reconnection has rather similar properties. Magnetic reconnection is stabilized by shear flow (Fig. 28). The reason for this is that the information velocity for magnetic reconnection is the Alfvén speed. Thus information that reconnection operates cannot propagate away from the x line if plasma flow is faster than the Alfvén speed such that reconnection cannot operate in the presence of fast plasma flow (La-Belle-Hamer et al., 1995 [43]; Chen et al., 1997 [24]). Therefore reconnection requires $\Delta V_A > \Delta v$ while the KH mode requires $\Delta v > V_{A,typ}$ along the k vector of the instability.

In the equatorial plane the geomagnetic field is strongly northward. Thus the KH instability can operate in the equatorial plane if the IMF is mostly northward or southward. Many observations show quasi-periodic signatures

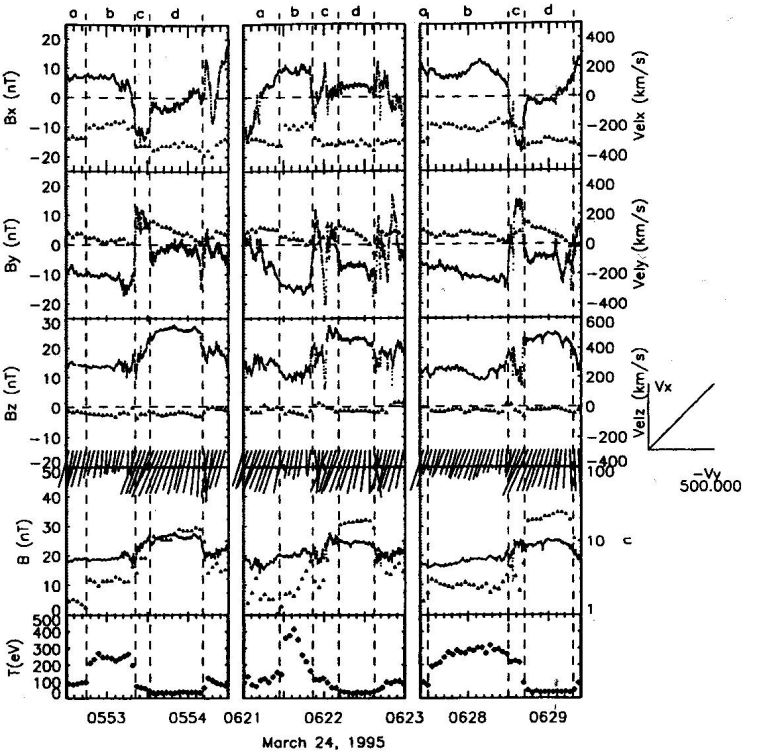


Fig. 29. Three examples of quasi-periodic plasma and field signature from Fairfield et al. (2000) [33].

of density, velocity and magnetic field fluctuations (Sckopke et al., 1981 [95]; Chen and Kivelson, 1993 [25]). A detailed study of such an event (Fairfield et al., 2000 [33]) demonstrated that these signature have a number of characteristic features (Fig. 29). One of the most remarkable feature is the transient presence of negative B_z components although the geomagnetic field and the IMF were strongly northward.

In comparison two-dimensional MHD simulations (Fig. 30) showed many of the same signatures and it was possible to identify individual signatures for instance for the entrance and exit of the SC into and out of the magnetosphere (Otto and Fairfield, 2000 [67]). Particularly the twisting of the magnetic field in the KH vortex motion can generate negative B_z signature in small regions of space.

Simulations also demonstrated another important property of magnetic field deformation in the KH vortex. The KH mode is an ideal instability and therefore does not permit mass transport across the magnetospheric boundary. As argued before the initial field is usually strongly parallel such that

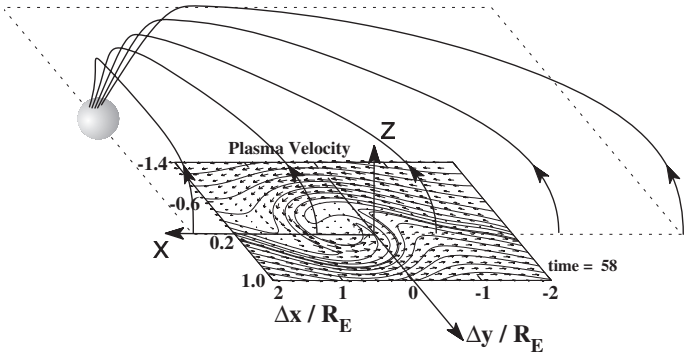


Fig. 30. Geometry for 2D MHD simulations at the magnetospheric boundary.

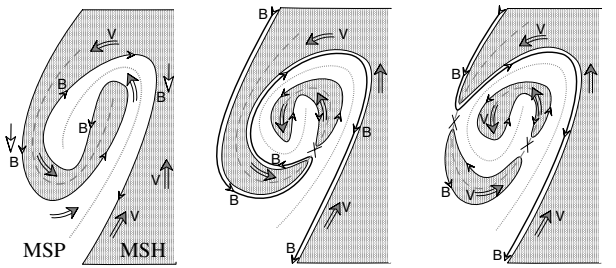


Fig. 31. Schematic of the magnetic field deformation and subsequent reconnection inside KH vortices.

reconnection cannot occur. However, the magnetic field deformation can generate strong local current sheets in the KH vortices (Fig. 31). Thus reconnection is possible within these small scale current sheets (Otto and Fairfield, 2000 [67]).

The mass transport into the magnetosphere for northward IMF is an unresolved problem. A possible mechanism for this transport is reconnection at high latitudes (above the northern cusp and below the southern cusp) which can connect IMF with the magnetosphere and thus capture magnetosheath material in the magnetosphere. An second plausible mechanism is reconnection inside KH vortices. In a quantitative evaluation of the reconnection inside of KH vortices (Fig. 32) it is shown that the transport rate is indeed sufficient to explain an effective mass diffusion coefficient of $D = 10^9 \mathrm{m}^2\mathrm{s}^{-1}$ (Nykyri and Otto, 2001 [59]).

It should be remarked that the KH mode may also occur at other locations on the magnetospheric boundary than discussed here. In fact for any orientation of the IMF there are always locations on the magnetospheric boundary where the magnetosheath and magnetospheric fields are approximately aligned.

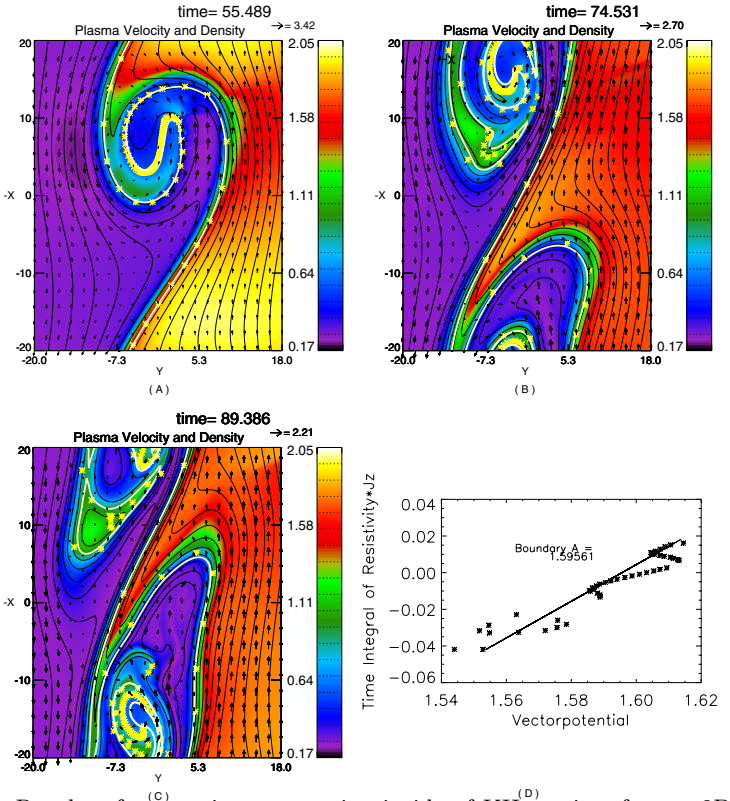


Fig. 32. Results of magnetic reconnection inside of KH vortices from a 2D MHD simulation.

Other Processes at the Magnetopause Boundary

Both magnetic reconnection and KH modes require an boundary that unstable with respect to either mode. However, energy and momentum can also be transferred by fluctuations carried by the solar wind or generated at the bow shock. It is suggested that solar wind pressure variations can also cause FTE signatures (Sibeck, 1990 [99]; Sibeck, 1992 [100]). While pressure variations can cause some of these signatures it is not expected that the corresponding signature would have similar statistical properties such as a correlation to southward IMF.

However, large increases in the solar wind velocity, density, or magnetic field have strong geomagnetic effects. They lead to a large scale compression of the magnetosphere and an intensification of all current systems. If a CME hits the magnetosphere the magnetopause can be pushed inside of 6 R_E which can for instance expose geostationary satellites to the solar wind.

It has also been suggested that plasma filaments with access momentum can actually penetrate the magnetopause and lead to a transport of mass into the magnetosphere (Lemaire, 1977 [48]). This process has been termed impulsive penetration. Analytic and numerical simulations indicate that this may be possible if the magnetosheath and magnetospheric fields are highly aligned (Schindler, 1979a [86]). However, the actual capture of the filaments inside the magnetosphere would still require magnetic reconnection. Thus far there is no observational evidence of impulsive penetration.

4 The Magnetotail

The basic elements of the magnetotail of the magnetosphere (Fig. 33) are

- Mantle (current): Region of open field with high density magnetosheath plasma.
- Lobes: Low density, strong magnetic field region. Energy is stored in the lobe magnetic field.
- Plasma sheet: Region of higher density and higher thermal pressure close to the equatorial plane. The plasma $\beta \geq O(1)$.
- Current sheet or neutral sheet : region of the cross tail current generating the lobe field. Here the magnetic field is rather weak thus the name neutral sheet.

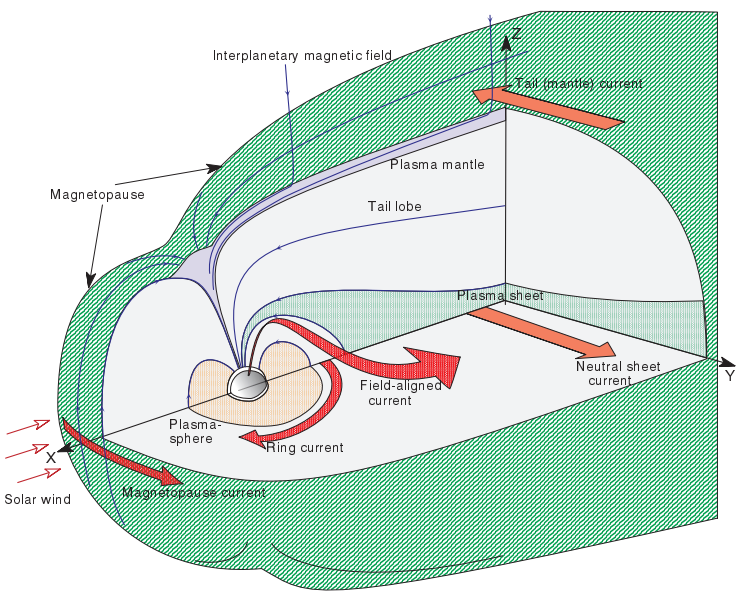


Fig. 33. Sketch of the magnetosphere with various elements of the magnetotail and the magnetospheric current systems.

– Field aligned currents: Birkeland currents originate from Earthward boundary of the current sheet and close in the ionosphere.

The magnetotail is the region of the magnetosphere where energy is stored. The tail plays an important role in magnetospheric substorms. Here the process of storage and release of energy in the tail determines much of the magnetospheric dynamics and is important for the coupling to the ionosphere and all related space weather effects.

4.1 Magnetotail Models

Magnetic Field Models

There are number of models which describe the magnetic field of the magnetosphere, e.g. Tsyganenko (1990) [111], Tsyganenko and Stern (1996) [113], and Tsyganenko (2000) [112]). These models use a suitable set of base functions and then fit the coefficients of these functions from large databases of satellite observations. The models can easily be parameterized for various IMF conditions, solar wind pressure, dipole tilt etc.

It is important to understand that these models are not equilibrium models. They do not provide any plasma data and in general it is not possible fit an isotropic pressure distribution to generate equilibria just from the magnetic field model.

Magnetic field models have many different application. For instance they can provide a reference for spacecraft observations. These models are frequently used to carry out test particle computations to study mechanisms which form distribution functions or to determine how particles can enter into certain regions of the magnetosphere. However, some caution is necessary particularly for this use. To compute particle trajectories both magnetic and electric fields are required. The latter are often assumed for instance as a constant in the cross tail direction. A constant electric field, however, implies $\partial \mathbf{B}/\partial t = 0$ and therefore implies stationary convection in the magnetotail. It is highly questionable whether such a steady state convection exists. This point is discussed later in this section.

While the magnetic field models are not equilibrium models they can be used to obtain typical properties of the magnetic field such as the magnetic flux tube volume (Fig. 34). This can for instance provide insight into magnetospheric convection.

Equilibrium Configuration

The magnetotail is surprisingly stable for long periods of time. Typically convection is small and the tail configuration is well described by equilibrium solutions. Analytic equilibria are available for the section of the magnetotail where the variation along the magnetotail (or in the cross-tail direction y) is small compared to the variation perpendicular to the current sheet (weakly

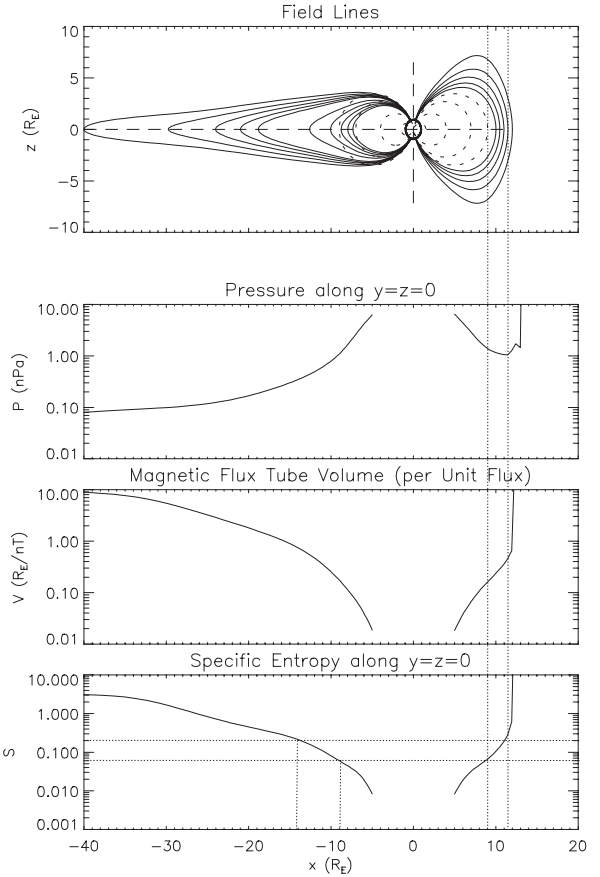


Fig. 34. Illustration of the magnetic field and some associated properties such as pressure, magnetic flux tube volume, and entropy on magnetic field lines.

two- or three-dimensional), e.g. Schindler (1975) [85], Schindler (1979a) [86], Birn et al. (1975) [14], and Birn et al. (1977) [15]. Because of the variation in x the solutions are applicable only at sufficient distances ($\geq 10R_E$) from the Earth. These equilibria can be constructed as fully kinetic solutions. In the MHD approximation they solve the static MHD equations.

A simple example for this class of analytic solutions is the following solution which represents a two-dimensional modification of the classic Harris sheet configurations (Harris, 1962 [37])

$$A_y = A_c \ln\cosh(z/l(x)) + f(x)$$
$$B_x = -\partial A_y/\partial z = B_0(x)\tanh(z/l(x))$$
$$B_z = \partial A_y/\partial x$$

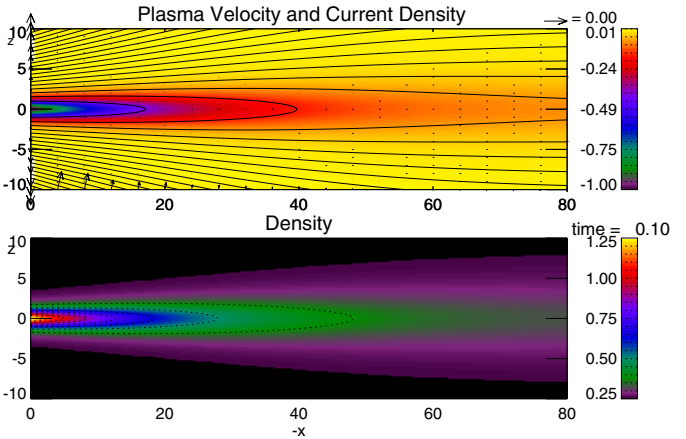


Fig. 35. Magnetic field and current density (top) and plasma density of a two-dimensional tail equilibrium.

Here A_y is the y component of the vector potential and $l(x)$ is a general function which can be used to match either the pressure or magnetic field variation along the magnetotail. Figure 35 shows a configuration with a realistic pressure variation along the tail. The Earthward boundary is located at approximately 15 R_E. This class of analytic solutions is particularly useful for a number of studies such as the stability of particular configurations, pressure anisotropy in the tail, and as initial configurations for numerical simulations.

In addition to analytic equilibrium solutions there are numerical equilibrium solutions for the magnetotail. The resulting configurations are not subject to the constraints of analytic equilibria and can be extended much closer to the Earth. They provide insight into the mechanisms which cause field-aligned current and the structure at the earthward edge of the plasma sheet. There are two basic approaches one of which solves the MHD equilibrium equations directly using numerical iteration methods. The other approach starts from a magnetic field model with a good guess for the plasma and pressure distribution. A numerical relaxation method is then used to obtain an equilibrium configuration.

Convection in the Magnetotail

During periods of southward IMF one can attempt to approximate convection with a constant cross-tail component of the electric field. Considering a cut in the noon midnight meridian it is possible to evaluate the velocity perpendicular to the magnetic field using the $\mathbf{E} \times \mathbf{B}$ drift (Fig. 36). For a constant electric field in the positive y direction (which should be equal to the dayside reconnection rate of$\approx 10^{-3}$ V/m) one obtains convection of lobe field lines (20 to 40 nT) toward the plasma sheet of a few 10 km/s. In the

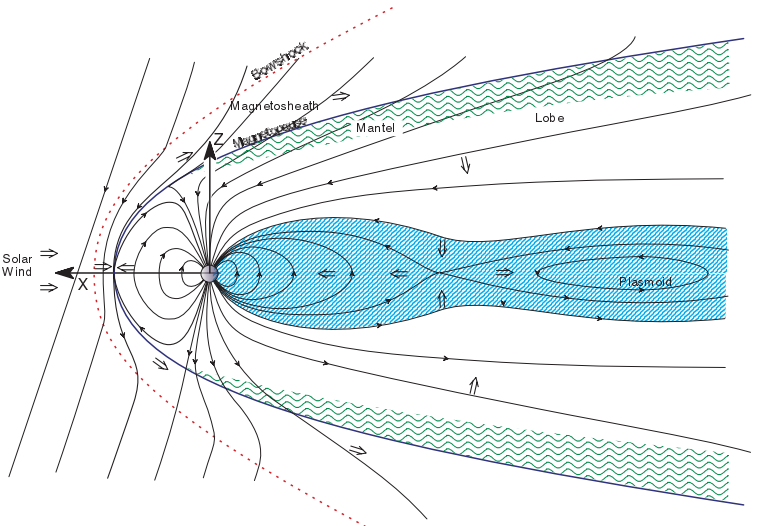


Fig. 36. Convection in the magnetosphere for southward IMF.

plasma sheet convection is directed toward the Earth with relatively large velocities of a few 100 km/s because of the weak magnetic field in the neutral sheet (few nT). The constant electric field implies $\partial \mathbf{B}/\partial t = 0$,i.e. stationary convection. Using long time averages (hours) the magnetic field may indeed not change much locally such that this pattern describes correctly the average convection over sufficiently long periods of time. However, there are important constraints to convection in the plasma sheet which need to be considered.

While stationary convection provide some insight into magnetospheric convection there are important constraints which prohibit stationary convection for most times. Using the MHD equations it is easy to show that the quantity $p\rho^{-\gamma}$ is conserved, i.e., $d\left(p\rho^{-\gamma}\right)/dt = 0$. Here $p\rho^{-\gamma}$ is a measure of entropy . By integrating over the length of a field line and defining the differential flux tube volume as

$$V = \int ds/B$$

one can find two additional conserved quantities which are the number of particles N and the specific entropy S on field lines.

$$N = \int \frac{\rho dl}{B}$$
$$S = pV^{\gamma}$$

Strictly these conservation laws apply only in ideal MHD and if there is no loss of particles or kinetic energy into the ionosphere. However, for typical

applications non MHD effects and/or loss in the ionosphere are negligible. The conservation laws imply that if a field line is convected from 40 R_E to 10 R_E the number of particles and the entropy remain constant. However, the field line volume is much larger for a typical field line originating at 40 R_E then it is at 10 R_E. Using the example of the Tsyganenko model (Fig. 34) the flux tube volume changes by about two orders of magnitude which implies that the pressure has to increase by more the 10^3 to maintain a constant entropy during convection. Note that the local magnetic field does not change for steady convection. However, this would yield an entirely unrealistic pressure close at 10 R_E such that convection for this field model cannot be stationary (Ericson and Wolf, 1980 [32]; Birn and Schindler, 1983 [11]; Schindler, 1979b [87]). Note that a realistic pressure distribution, which yields approximately an equilibrium in the noon midnight meridian, yields an entropy that varies by two orders of magnitude (Fig. 34) thereby excluding steady state convection. Both density and specific entropy distributions from satellite observations are not consistent with steady state convection from the mid-tail (30 to 40 R_E) to the near Earth region.

Several model describe a quasistatic (slow) evolution of the magnetotail (e.g. Schindler, 1979b [87];Schindler and Birn, 1982 [89]) and demonstrate that the electric field (or convection) is shielded from the plasma sheet.

4.2 Magnetospheric Substorms

Magnetospheric substorm are major events in the magnetosphere. During a substorm a large amount of energy is first stored in the magnetosphere with a subsequent fast release of this energy. A typical substorm consists of three distinct phases: A growth phase during which energy is accumulated mostly in the magnetotail. An expansion phase during which energy is released, and a recovery phase during which the magnetosphere returns mostly to its original state.

Growth Phase

Typical for almost all substorms is an initial southward turning of the IMF. Due to the southward IMF there is enhanced reconnection on the dayside magnetopause and magnetic flux is removed from the dayside and accumulated in the lobes of the magnetosphere. The physical consequences of this process are well documented. During the growth phase the size of the polar cap (region of open field lines in the ionosphere) increases gradually. Radar observations show that convection in the polar cap increases. The cross polar cap potential increases, which is an measure for of global reconnection rate and of the transport of magnetic flux from the dayside to the nightside. Satellite observations show that the magnitude of the lobe magnetic field strength increases and the magnetotail configuration stretches in particular in the near Earth tail where the field is usually more dipolar.

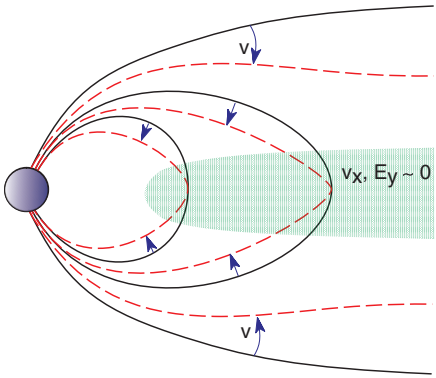


Fig. 37. Sketch of the reconfiguration (red lines) of the near Earth magnetotail.

A particularly important property is the thinning of the near Earth current sheet inside of 15 R_E. Satellite observations show that the magnetic geometry of this region is changing from dipolar to a stretched tail-like configuration as illustrated in Fig. 37. The thickness of the current sheet in this region (and most pronounced at about 10 R_E) decreases gradually (Sergeev et al., 1990 [97]; Pulkkinen et al., 1994 [79]). Some observations report an explosive thinning during the last minutes of the growth phase (Ohtani et al., 1992 [60]) meaning that during this time the thickness decreases rapidly. The reported current sheets are often only 500 to about 1000 km thick representing a collapse of the current sheet to less than a tenth of its original width. In the midnight ionosphere the distance between the most equatorward discrete arc and the diffuse Aurora (at the equator boundary of the auroral oval) decreases.

The current sheet thinning is of major importance for the subsequent evolution because it provides the conditions for the instability that leads into the growth phase. As pointed out above the onset of dayside reconnection cannot simply generate stationary convection. Note that simple compression due to enhanced lobe field can reduce the current sheet thickness only by a rather small amount for realistic parameters. One model for the current sheet thinning suggests that microturbulence generates a diffusion of the thermal pressure (Lee et al., 1998 [47]). This pressure diffusion implies for the equilibrium a different pressure distribution on magnetic flux tubes $p\,(A_y)$ and since the equilibrium current density is $\sim dp/dA_y$ this can have profound implications for the evolution of the current sheet. Another very promising model using quasistatic models finds that even small boundary perturbations can change the resulting equilibrium properties considerably up to the extend that equilibria seize to exist for some perturbations (Schindler and Birn, 1989 [90]; Schindler and Birn, 2002 [91]; Birn and Schindler, 2002 [12]). According to this model the formation and strength of current sheets in the magnetotail

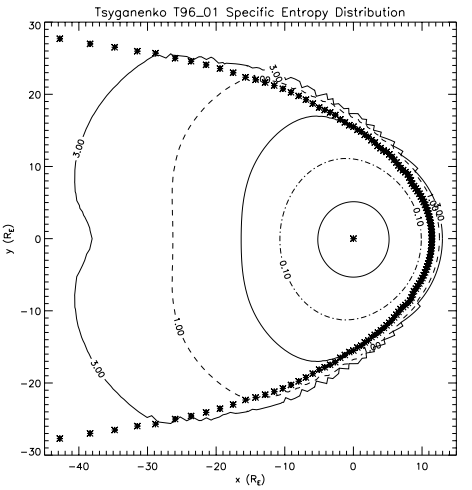


Fig. 38. Distribution of specific entropy mapped into the equatorial plane.

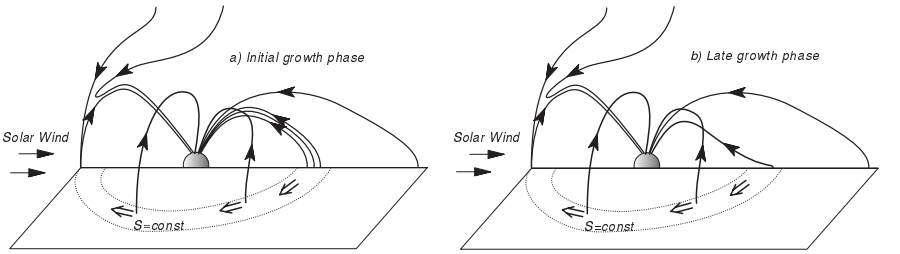


Fig. 39. Sketch of convection from the tail to the dayside magnetopause.

does not only depend on the magnitude of the perturbations but also on the temporal history of how the perturbation is applied (Birn et al., 2003 [13]).

The specific entropy provides another clue for the current sheet thinning. Figure 38 shows a map of the specific entropy obtained from a Tsyganenko magnetic field model. Considering the removal of magnetic flux on the dayside this flux has to be replaced from the nightside. The entropy map shows that convection from a region at about 10 R_E in the tail can replace magnetic flux which has been removed by reconnection from the dayside magnetosphere. Thus the the near Earth tail represents a magnetic flux reservoir from which flux is removed during the growth phase as illustrated in Fig. 39. It is conceivable that this mechanism contributes to the current sheet thinning because it can explain the location of the current sheet thinning and the approximate duration of the growth phase.

Expansion Phase

The expansion phase of a substorm is the phase in which previously stored energy is released (Akasofu, 1964 [1]; McPherron, 1979 [52]). The name expansion phases originates from auroral observations. In these observations the most equatorward discrete arc starts to brighten and the auroral activity is expanding rapidly poleward and westward (westward traveling surge). The duration of the expansion phases is typically 30 to 60 minutes, however, with strong activity for about 15 minutes. Further ionospheric features are a strong increase in the auroral electrojet (a strong hall current in the ionosphere which provides the main magnetic perturbations measured on the ground) and a correspondingly strong intensification of the Birkeland currents. These are field-aligned currents which are essential for the coupling between the ionosphere and the magnetosphere.

Satellite observations in the magnetotail show a rapid dipolarization (return to a dipolar magnetic field configuration) in the near Earth magnetotail. The dipolarization is often explained by a disruption of the cross-tail current, however, one should keep in mind that current and magnetic field are equivalent and one should not interpret the local disappearance of the cross-tail current as a physical mechanism for the reformation of the dipolar field. Geosynchronous satellites measure significant increases in the energetic particle fluxes (particle injection). In the mid and far tail satellites observe signature of large plasma bulges (plasmoids) traveling down-tail. These plasmoids are accompanied by fast tailward plasma jetting down-tail. Satellites closer to Earth see similar plasma jetting, however, usually toward the Earth. Finally satellites in the lobes observe frequently a transient compression (travelling compression region or TCR) of the lobe magnetic field and a subsequent decrease of the lobe field strength.

In recent years much discussion focused specifically on the onset mechanism. Earliest signatures of the onset include the brightening of the onset arc, dipolarization, particle injection, and fast plasma flows. The appearance of these signatures can be measured in some cases within one minute of each other and any one of the signatures can be the first. A precise timing of a particular sequence of these signatures proved rather difficult because they occur locally and spread over time such that a spacecraft has to be exactly in the correct location to observe the earliest signature.

The mechanism for the onset of the expansion phase is not known. For a long time the ion tearing mode was a major candidate for the onset. However, in a thick current sheet with a normal magnetic field component this mode is strongly stabilized. This changes in the light of strong current sheet thinning. However, it could also be that the current sheet formation leads to a loss of the equilibrium or that the thin current sheets are unstable to either microinstabilities or to shear-flow (KH type) instabilities. All of these mechanism can more easily operate in a thin current sheet. There is evidence that a significant fraction of onsets is triggered by northward turnings of the

IMF. However, it is yet not clear whether the resulting change in convection contributes to the trigger or whether the northward turning just provides a nonlinear perturbation for a marginally unstable current sheet.

Despite the discussion on the onset mechanism the subsequent events are mostly well explained by magnetic reconnection as illustrated by the simulation results in Fig. 36 (Birn, 1980 [8]; Birn and Hones, 1981 [10]; Otto et al., 1990 [69]; Hesse and Birn, 1991 [38]. The simulation is carried out for the equilibrium shown in Fig. 35. Figure 40 shows the onset of reconnection and the formation of a plasmoid which is ejected tailward. The plasmoid has a core of enhanced plasma density and pressure. While the plasmoid is traveling down-tail the lobe field is temporarily compressed and after the passage of the plasmoid it settles back to a smaller than initial value because the plasmoid has removed some of the magnetic flux. This explains the transient amplification of the lobe field as observed for TCR's. Fast tailward plasma flow is observed down the tail while there is some enhanced Earthward flow in the near Earth region. The magnetic field close to the Earthward boundary becomes much more dipolar (large z component) because reconnection increases the flux through the equatorial plane Earthward of the x line. Finally the appearance of energetic particle fluxes at geostationary orbits is due to adiabatic particle motion. During the growth phase the magnetic field was relatively weak in this region. Particles on reconnected field lines convect toward the Earth into a region of much stronger magnetic field. The conservation of the magnetic moment $\mu = mv_\perp^2/2B$ for these particles implies that the perpendicular energy can increase by several orders of magnitude depending on the original and final field strength. This process is called betatron acceleration and can indeed explain the observed increase in energetic particle in the near Earth region (Birn et al., 1997 [16]).

It should also be noted that magnetic reconnection not only can explain these various observations but that it is required in the expansion phase. During the growth phase a large amount of closed magnetic flux has been opened. The only mechanism by which this flux can be closed again is reconnection in the magnetotail. Finally magnetic reconnection also solves the problem of entropy conservation during convection because entropy conservation only applies to ideal MHD (Hall MHD) and reconnection changes the entropy distribution because field lines are newly connected.

Recovery Phase and Other Substorm Related Phenomena

The recovery phase completes the substorm sequence and is the return of the magnetospheric configuration to its original configuration. The duration is typically 1 to 2 hours. During the recovery phase the auroral oval is dimming and the aurora moves back to higher latitudes. Typical are pulsating patches of auroral and quiet arcs reappear. Similarly the current sheet and plasma sheet returns to its original size.

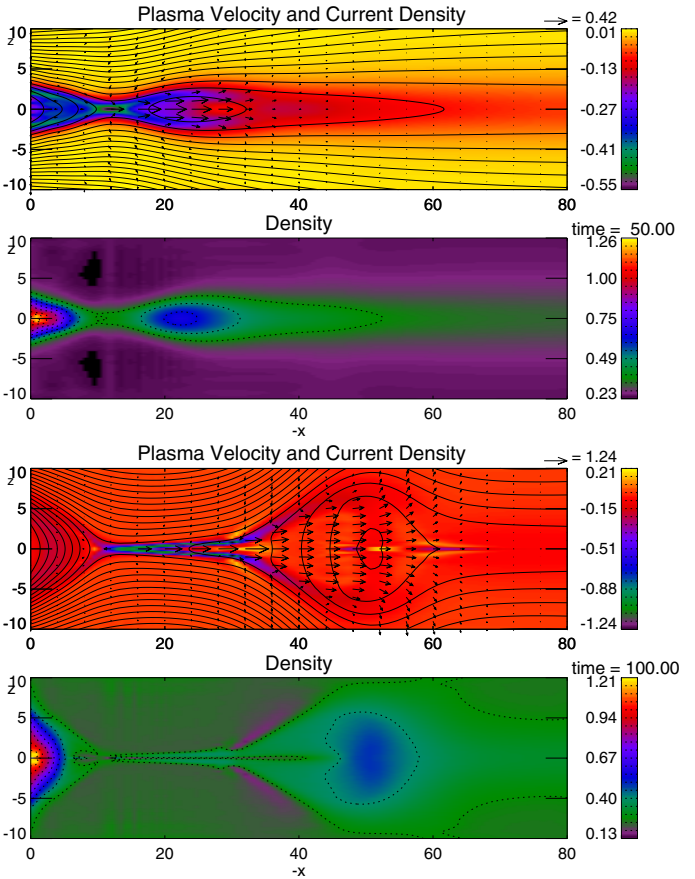


Fig. 40. Results from a two-dimensional MHD simulation of reconnection in the magnetotail.

This does not mean that the recovery phase is entirely understood. An important unresolved question is the origin of the plasma sheet material or in other words how the plasma sheet is reformed.

It should also be mentioned that the recovery phase can be rather brief or entirely absent. During prolonged periods (many hours) of southward IMF with many substorms there is often not a complete return to a pre-substorm state of the magnetosphere. Similarly there may not be a clearly identifiable growth phase during such times.

In particular during prolonged periods of relatively steady southward IMF there is a state of the magnetosphere which is termed steady magnetospheric convection events (SMC's) (Sergeev et al., 1996 [96]). Note that this term should not be taken literally in the mathematical sense. During SMC's there is continuous reconnection at the dayside magnetopause. Typical for SMC's are

so-called bursty bulk flows (BBF's) (Angelopulos et al., 1992 [4]). These are events which show several minutes of very fast plasma flow which can be tail- or Earthward. Bursty bulk flows can actually occur all times, however, usually rather infrequent. During SMC's, BBF's are very frequent, the Aurora is highly active, and the auroral electrojet is strong. Intuitively it is conceivable that BBF's are localized reconnection events in the magnetotail in particular because magnetic flux must be re-closed in the tail and transported back to the dayside. Although there is considerable auroral activity the large scale organized release of energy as observed in substorms is absent for SMC's. Thus the magnetosphere has two different ways to respond to the same IMF and solar wind conditions. Why it would respond one way or the other is unresolved.

4.3 Magnetosphere – Ionosphere Coupling

The coupling between the magnetosphere and the ionosphere is important because (1) there is a large amount of energy deposited in the ionosphere and (2) the ionosphere provides in addition to the solar wind the only boundary to the magnetosphere.

Traditionally magnetosphere - ionosphere coupling assumes a stationary state. In this case $\partial \mathbf{B}/\partial t = \nabla \times \mathbf{E} = 0$ implies that the electric field can be derived from a potential $\mathbf{E} = -\nabla\phi$ and for $\mathbf{E} \cdot \mathbf{B} = 0$ this potential is constant on magnetic field lines and maps into the ionosphere. Because of the large phase velocity of the fast wave compressibility can be neglected for ionospheric convection. Since the electric field is caused by convection, contour lines of the potential are flow lines in the ionosphere.

This simple model has been relatively successful. For instance radar observation of the ionospheric flow show the enhancement of the polar cap potential during periods of southward IMF and are used to study the transport of magnetic flux in the magnetosphere. Similarly radar observations on the nightside have provided interesting new insight into the convection during the expansion phase. Until recently it was believed that ionospheric convection increases at expansion phase onset because fast flow and enhanced convection is observed in the magnetotail. However, Superdarn radar observation show actually the opposite. Convection appears to slow down at onset and increases gradually over the course of about 10 minutes to about the growth phase level (Bristow et al., 2001 [18]; Bristow et al., 2003 [19]).

This observation is remarkable because it is either inconsistent with the assumption of a steady state or it implies that the satellite observations in the magnetotail are incorrect. The solution to the inconsistency is simple. The large scale magnetosphere is not in a steady state. Therefore the assumption that the electric field maps into the ionosphere is incorrect. During the growth phase convection continues in the ionosphere in the midnight sector although convection in the plasma sheet is slow. This is not surprising because the ionospheric flow is incompressible and therefore has to continue somehow.

Finally it is worth to note that the ionosphere can have significant influence on magnetospheric dynamics. Strong ionization (due to precipitation) can increase the ionospheric conductivity bu orders of magnitude. The enhanced conductance causes an almost ideal reflection of Alfvén waves which imposes a strong friction on the magnetospheric flow (Kan, 1998 [39]). Similarly it is now well established that ion outflow of heavy ions from the ionosphere can exert considerable drag in the magnetosphere.

5 The Inner Magnetosphere – Geomagnetic Storms

The inner magnetosphere plays an important role in magnetospheric dynamics in particular for magnetic storms (e.g. Wolf, 1995 [116]). During storms the ring current (Fig. 33) becomes very strong and this current generates strong perturbations of the magnetic field on the ground particularly at low latitudes. The signatures of selected low latitude magnetometer stations are combined to the so-called Dst magnetic index. The ring current is generated by energetic particle populations in the inner magnetosphere. To understand the physical mechanisms that generate the ring current it is necessary to understand the dynamics of trapped particle in the inner magnetosphere.

The inner magnetospheric dynamics is considerable different from the physics in the outer magnetosphere. The dipole magnetic field is much stronger than the magnetic field perturbations. Thus it is mostly sufficient to neglected the magnetic field perturbations in the inner magnetosphere and to analyze only the particle dynamics. The basic particle motion is a combination of particle drifts and bounce motion in the geomagnetic field.

5.1 Magnetic Field and Basic Particle Properties

Magnetic Field

As outlined above the magnetic field can be described to lowest order in terms of the dipole field (Fig. 41), e.g. Baumjohann and Treumann (1997)

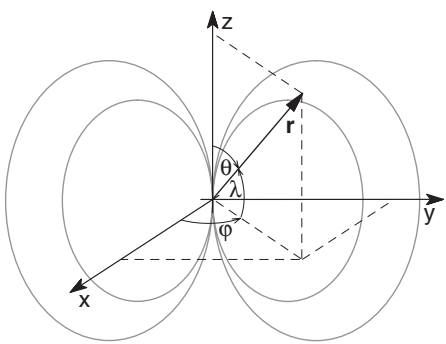


Fig. 41. Dipole geometry.

[6]. Using the latitude λ the dipole field components are given by

$$B_r = -2B_E \frac{R_E^3}{r^3} \sin\lambda \tag{14}$$

$$B_\lambda = B_E \frac{R_E^3}{r^3} \cos\lambda \tag{15}$$

where $B_E = \mu_0 M_E / 4\pi R_E^3 = 3.11 \cdot 10^{-5}$ T is magnetic field at the equator on the Earth's surface.

To evaluate magnetic field properties is is helpful to evaluate the magnetic field magnitude

$$B = B_E \frac{R_E^3}{r^3} \left(1 + 3\sin^2\lambda\right)^{1/2}$$

and to use an equation to parameterize magnetic field lines $\widetilde{r} = L\cos^2\lambda$ where L is the distance (in r_E) at which a field line crosses the equatorial plane (McIlwain L parameter) and $\widetilde{r}$ is measured in Earth radii r_E. This parametrization allows for instance to determine the magnetic field magnitude along a specific field line just as a function of the latitude $B = B_E \left(1 + 3\sin^2\lambda\right)^{1/2} / \left(L^3 \cos^6\lambda\right)$.

Particle Motion

Mirror Motion: For periodic particle motion with a period smaller than changes of the overall system the action integral is an approximate constant of motion and called an adiabatic invariant.

$$J_i = \oint p_i dq_i = const \tag{16}$$

This yields the magnetic moment $\mu = W_\perp / B$ as the first adiabatic invariant for the gyro motion of charged particles where $W_\perp = m v_\perp^2 / 2$ is the particle energy based on the velocity perpendicular to the magnetic field. Defining the particle pitch angle α as the angle between the particle velocity and the direction of the magnetic field such that $v_\perp = v \sin\alpha$ the magnetic moment becomes

$$\mu = \frac{m v^2 \sin^2\alpha}{2B} \tag{17}$$

which implies $\sin^2\alpha / B = \text{const}$. A particle is mirrored (reflected) in the magnetic field (assuming that the electric field is negligible) when the pitch angle assumes 0. Thus a particle with a pitch angle α_{eq} in the equatorial plane is mirrored at the point where the magnetic field along the field line is $B_m = B_{eq} / \sin^2\alpha_{eq}$ as illustrated in Fig. 42.

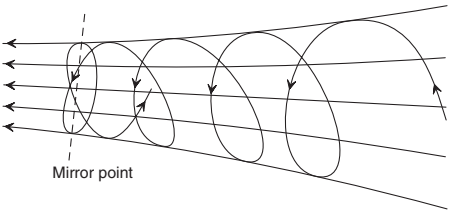


Fig. 42. Illustration of magnetic mirror motion.

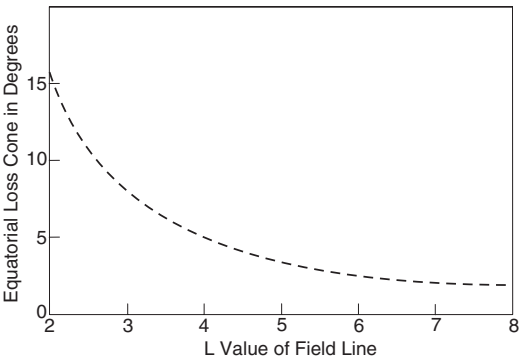


Fig. 43. Equatorial loss cone as a function of L value.

This provides another important property of the particle distributions. If a particle is not mirrored above the ionosphere it will be lost due to collisions in the neutral atmosphere. This defines a critical angle α_l in the velocity distribution for which all particles with a pitch angle smaller than α_l (with $\sin\alpha_l = \sqrt{B_{eq}/B_0}$ where B_0 is the magnetic field on the ground) are lost from the particle distribution. Using the magnetic field relation one can express the equatorial loss cone directly as a function of the L value $\sin^2\alpha_l = \left(4L^6 - 3L^5\right)^{-1/2}$ which is shown in Fig. 43. The loss cone becomes larger with distance from the equatorial plane or higher latitudes. Since there are very few collisions the loss cone is usually almost empty.

Using the magnetic field equations we can also derive a relation between the latitude λ_m of the mirror points and the equatorial pitch angle. This allows to determine the bounce period for the mirror motion as $\tau_b = 4\int_0^{\lambda_m} ds/v_\parallel$. The integral can be evaluated to obtain bounce period as function of particle energy, equatorial pitch angle, and L value of the field line.

$$\tau_b \approx \frac{LR_E}{(W/m)^{1/2}}\left(3.7 - 1.6\sin\alpha_{eq}\right) \tag{18}$$

Figure 44 show the bounce period for 1 keV electrons and ions with an equatorial pitch angle of 30°.

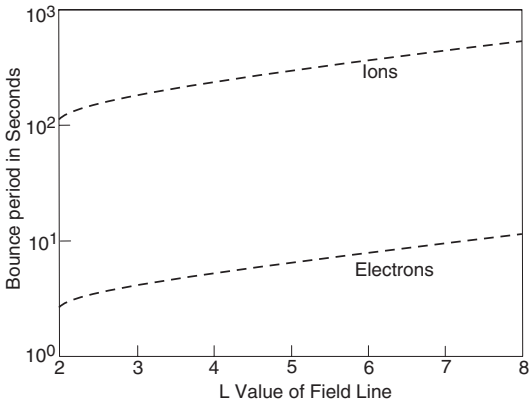


Fig. 44. Bounce period as a function of L value for 1 keV particles and $\alpha_{eq} = 30$.

Drift Motion: In addition to the gyro and the bounce motion there are various drifts which are important for the particle dynamics. It is in fact the drift motion which generates the ring current. Applying a general force $\mathbf{F}$ perpendicular to the magnetic field generates the general drift $\mathbf{v}_F = \mathbf{F} \times \mathbf{B}/(qB^2)$. There are three basic particle drifts, the electric force or $\mathbf{E} \times \mathbf{B}$ drift, the magnetic gradient drift, and the magnetic curvature drift.

$$\mathbf{v}_E = \frac{\mathbf{E} \times \mathbf{B}}{B^2}$$
$$\mathbf{v}_\nabla = \frac{mv_\perp^2}{2qB^3} (\mathbf{B} \times \nabla B)$$
$$\mathbf{v}_C = \frac{mv_\parallel^2}{qB^4} \mathbf{B} \times [(\mathbf{B} \cdot \nabla) \mathbf{B}]$$

Here the magnetic gradient drift is caused by the force exerted on a particle in the presence of a gradient in the magnetic field and the curvature drift is caused by the centrifugal force on a particle moving along a curved magnetic field line. While the $\mathbf{E} \times \mathbf{B}$ drift is the same for electrons and ions and therefore does not generate a current, the two other drifts depend on the charge and the perpendicular respectively parallel particle energy such that these drifts give rise to a nonzero current density. Note that these relations are valid only if the gradients and curvature terms are small compared to the the gyro radius effects and if the resulting current do not modify the magnetic field too much.

During the bounce motion a particle will also undergo gradient and curvature drift. The instantaneous angular drift velocity in the is $v_d/r \cos\lambda$. By integrating this over the bounce period one can determine the average angular drift $\langle \omega_d \rangle = \Delta\psi/2\pi\tau_b$ or the drift period $1/\langle \omega_d \rangle$ for particle to encircle

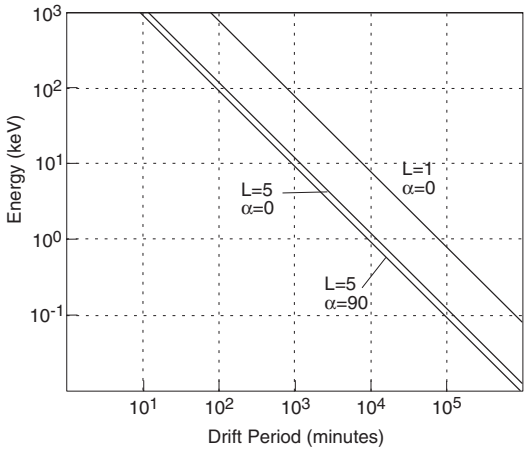


Fig. 45. Drift period in the Earth's dipole field.

the Earth

$$\tau_d \approx \frac{2\pi q B_E R_E^2}{3WL} \left(0.35 + 0.15 \sin \alpha_{eq}\right)^{-1} \tag{19}$$

with $W = mv^2/2$. Figure 45 shows that 1 keV particle has drift periods of 100 to several hundred hours while a 1 MeV particle has drift periods of 10 to 100 minutes. It is interesting to note that drift periods are shorter for larger L values. The energy dependence leads to a separation of particles with different energies if they are injected at the same location. Thus a satellite would observe a time lack (dispersion) for the arrival of populations with different energies which is used to identify the origin of the particle injection. Note that different pitch angles can also cause a separation of particles from the same origin.

5.2 Plasmapause, Alfvén Layer, and Ring Current

In addition to the gradient curvature drift there are two other particle drifts which need to be considered for the particle motion in the inner magnetosphere. The first of these is connected with the corotation of the magnetic field. This drift can be expressed as $\mathbf{v}_E = \omega_E r \mathbf{e}_\varphi$ with the angular velocity of the Earth's rotation ω_E and the eastward unit vector $\mathbf{e}_\varphi$. The corresponding electric field $\mathbf{E} = -\mathbf{v}_E \times \mathbf{B}$ in the equatorial plane can be expressed in terms of the corotation potential $\Phi_{cor} = -\omega_E B_E R_E^3/r$. This potential implies that particles are rotating with the Earth. The second drift is given by the plasma convection. For simplicity we assume a constant electric field E_0 across the magnetosphere. While this is not entirely realistic it provides insight into the basic effects of convection in the equatorial plane. Using the angle φ measured from the sun-Earth line in the Eastward direction on the dayside, the corresponding potential is $\Phi_{conv} = -E_0 r \sin \varphi$. Finally we will confine the

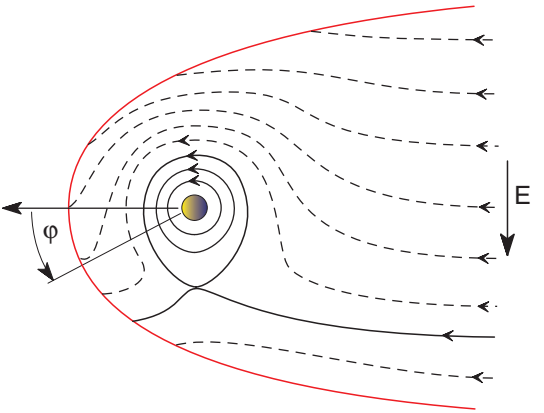


Fig. 46. Low energy particle drifts in the equatorial magnetosphere.

motion to the equatorial plane, i.e., consider only particles with 90° pitch angle. The corresponding gradient B drift potential is $\Phi_\nabla = \mu B_E R_E^3 / \left(qr^3\right)$. The resulting total drift velocity is

$$\mathbf{v}_D = \frac{1}{B^2}\mathbf{B} \times \nabla \Phi_{tot}$$

with $\Phi_{tot} = \Phi_{conv} + \Phi_{cor} + \Phi_\nabla$. The gradient drift is small compared to the corotation for sufficiently low particle energies (Fig. 45). Thus the net particle motion is given by the sum of the corotation potential and the convective potential $\Phi_{pp} = \Phi_{conv} + \Phi_{cor}$. The resulting drift paths (contours of constant Φ) are sketched in Fig. 46.

Close to the earth the drift paths are circular while they become almost straight lines at large distances. The direction of corotation and convection are opposite to each other on the duskside which generates the bulge in the closed (trapped) drift region. The location of zero velocity is $r_{pp}^2 = \omega_E B_E R_E^3 / E_0$. Thus the region of trapped particles shrinks in size for large values of the convection electric field. The region inside the separatrix represents the plasmasphere, a region with relatively cold plasma confined to a few R_E around the Earth. The separatrix or boundary of the trapped low energy population is called the plasmapause. Observations show that the plasmasphere indeed shrinks in size during times of magnetic activity and satellite observations demonstrate the disappearance of much of the plasmasphere during strong substorms.

The motion of energetic particle can be constructed in a similar way. For these particles one can neglect the corotation potential in favor of the gradient drift potential. The resulting motion is given by $\Phi_{rc} = \Phi_{conv} + \Phi_\nabla$ and the drift paths are sketched in Fig. 47.

Energetic particles are trapped closer to the Earth inside the separatrix (in red). The radial distance for the zero flow condition is $r_{rc} =$

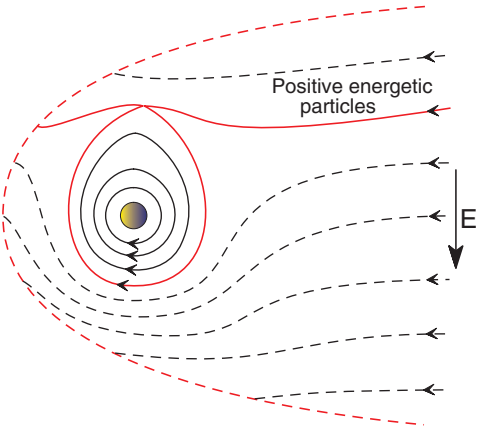


Fig. 47. Energetic ion drift paths in the equatorial plane.

$3\left(\mu B_E R_E^3/\left|q\right| E_0\right)^{1/4}$. As in the case of low energy particles the trapped region shrinks for strong convection which dominates in the outer magnetosphere. However, different from the case of low energy particles, the size of the trapping region increases with particle energy. Note that the drift is charge dependent such that the corresponding configuration for energetic electrons is mirrored across the noon midnight meridian. The separatrix between trapped and non-trapped particles is called the Alfvén layer.

Ring Current: The drift current for particles with an equatorial pitch angle of $\alpha_{eq} = 90°$, energy W, and density n on a given L shell is

$$j_d = \frac{3L^2 nW}{2B_E R_E}$$

which represents a current in the westward direction. Integrating this current over energy yields a total current of $I_L = 3U_L L/\left(2\pi B_E R_E^2\right)$ where U_L is the total particle energy at the distance $r = LR_E$. By including the diamagnetic effects due to gyro motion of all charged particles the magnetic perturbation at the Earth's center can be computed from Biot-Savarts law. With the total energy of all drifting particles U_R the combined disturbance from the particle drifts and the diamagnetic effect becomes

$$\Delta B = -\frac{\mu_0}{2\pi}\frac{U_R}{B_E R_E^3}.$$

Noting that the total energy contained in the dipole field is $U_{mag} = 4\pi B_E^2 R_E^3/\left(3\mu_0\right)$, the magnetic field disturbance becomes

$$\frac{\Delta B}{B} = -\frac{2}{3}\frac{U_R}{U_{mag}}$$

Thus the ground magnetic signature at low latitudes provide direct insight into the strength of the ring current and the total energy of the particles which cause the disturbance. For follow up discussions and further reading we refer to Parker (1991) [72], Roederer (1970) [80], Wolf (1995) [116].

5.3 Magnetic Storms

Large solar flares or CME's lead to strong perturbations of the terrestrial magnetosphere for long periods of time (days). Typically the dynamic solar wind pressure is strongly enhanced for many hours leading to a considerable compression of the entire magnetosphere. The effects of a CME can be amplified if in addition the IMF is very large and southward. In this cases it is possible to have several substorms over the course of a magnetic storm. Typically electric fields in the magnetosphere are also strongly enhanced.

The disturbance of the magnetosphere is measured as a strong perturbation of the horizontal ground magnetic field particularly at low (equatorial) geomagnetic latitudes. Selected magnetometer measurements are combined to determine the Dst index to provide an average measure of the magnetic disturbance. An example of the Dst signature of a magnetic storm is shown in Fig. 48. After an initial brief increase the Dst index can drop by several 100 nT within a few hours.

Magnetic storms have to phases. First one observes a strong decrease of the Dst index over the course of several hours and the Dst index stays low for about 10 to 20 hours. Subsequently there is a second phase of a slow (days) recovery of the Dst index.

As outlined above the cause for the strong magnetic disturbance is a large increase in the ring current energy. Assuming that at least half of the depression during storms is caused by the intensification of the ring current, yields

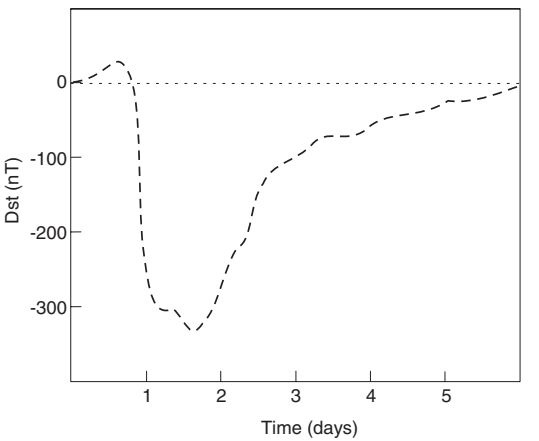


Fig. 48. Equatorial magnetic field depression (Dst) during a magnetic storm.

an estimate of the associated total particle energy $\Delta U_R \geq (\pi/\mu_0) B_E R_E^3 \, Dst$. An increase of $2 \cdot 10^{13}$ J in ring current energy corresponds to 1 nT perturbation of the ground magnetic field. Thus large magnetic storms can generate an increase of the ring current energy by more than $5 \cdot 10^{15} J$ and cause a total current of more than 10^7 A. This corresponds to disturbances of more than 1 % of the total dipole magnetic field energy. With an average energy of 10 keV per particle more than $3 \cdot 10^{30}$ particles are injected into the ring current at such times.

Sources and Sinks of Ring Current Particles

Source Mechanisms: The relative importance of various source mechanisms is not fully understood. Our prior discussion indicates that at least some of the ring current particles are injected from the plasma sheet due to reconnection. For these particle adiabatic heating (betatron effect) is one of the main acceleration mechanisms. For instance a particle starting from 8 R_E down-tail doubles it energy by drifting to 6 R_E and increases it energy by a factor of 20 when it drifts to $L = 3$. Note that typical plasma sheet ions start with energies in the few keV range. In addition to the conservation of the magnetic moment particles are also accelerated by the bounce motion (because they move into a region with decreasing length of the field lines). The different perpendicular and parallel heating mechanism yield anisotropic particle distributions.

However, it is documented that particle are also locally heated. One mechanism is the presence of very large perpendicular electric fields in the inner magnetosphere during the first phase of magnetic storms. In addition particles can be accelerated by turbulence and strong wave fields.

Observations also show that there is a significant portion of heavy ion (oxygen, etc.) present in the energetic population. These particles can only be of ionospheric origin or may have been particles which were initially present in the plasmasphere such that they must have been locally accelerated.

Loss Processes: Similar to the source processes, loss mechanisms are still researched. Among the more important loss processes are the following. Charge exchange collisions with neutrals extracts energy out of the energetic particles. Basically the process generates an energetic neutral particle which is not anymore bound by the magnetic field and a low energy ion. Pitch angle diffusion can also be caused by collisions or by plasma turbulence. The process brings energetic particle into the loss cone such that they are subsequently lost in the atmosphere. Varying electric fields and collision processes can move particle across the Alfvén layer (radial diffusion) such that they are not anymore trapped. Interchange motion can move flux tubes also radially in or outward such that their orbits are not anymore closed. The various processes are relatively complex and they depend on particle energies, compositions, presence of turbulence etc. Currently these processes are being studied with

large scale numerical simulations which follow different particle species and solve convection and diffusion processes in velocity space.

For many more details both on the solar origins and the inner magnetospheric processes we refer to Parker (1991) [72], Tsurutani et al. (1997) [110], Wolf (1995) [116].

6 Conclusions

While the basic formation of the magnetosphere is a very simple boundary problem (magnetic dipole exposed to the solar wind) the detailed structure and dynamics of the magnetosphere is rich in interesting physics.

The term space weather implies – similar to ordinary weather – environmental conditions which effect human life or installations. On the positive side magnetospheric dynamics can generate beautiful displays of Aurora which have fascinated humankind for ages. However, there are many space weather effects which are hazardous to human installations on the ground and in space.

Geomagnetic activity can induce electric fields and currents on the ground which may cause enhanced corrosion in pipelines and failure of power lines and transformers. Enhanced radiation is a potential problem in high altitude flight and similarly energetic particles in the radiation belts and ring current cause significant radiation problems for astronauts. Intense radiation and large electric fields may also have adverse effects on space instrumentation and cause the failure of satellites. Ionospheric modifications can interfere with radio communication and last not least the energy deposited in the high altitude atmosphere can cause a more expanded atmosphere with enhanced friction and reduced lifetimes for low orbiting satellites.

In some cases it may be sufficient to understand these hazards better to improve on instrumentation or to provide appropriate protection for astronauts. For many situations with a potential hazardous impact it is desirable to provide better forecasting of geomagnetic events. For both purposes it is important to understand the relevant physics in the different regions of the magnetosphere.

There is an increasing number of global magnetospheric simulations which attempt the forecasting. To some degree these global simulation methods are very successful. For instance they can reasonably well predict overall changes on large scales in the magnetosphere. For instance certain aspects of the large scale storm dynamics is rather well captured by global models. However, there are also a number of fundamental questions regarding the possibility of forecasting space weather. Similar to ordinary weather the magnetosphere may well be chaotic and it is not clear for instance whether or not substorms can in principle be predicted with sufficient accuracy.

This article attempted to summarize the basic dynamics of typical magnetospheric phenomena and to shed light on the important and interesting

physical mechanisms which cause magnetospheric activity. The field of magnetospheric research has made enormous progress in the past decades. The progress is valuable not only for space weather related issues but for many space plasma systems.

References

1. S.-I. Akasofu: *Planet. Space Sci.*, **12**, 273, 1964.
2. H. Alfvén: *Tellus*, **9**, 92–6, 1957.
3. B. J. Anderson, T.-D. Phan, and S. A. Fuselier: *J. Geophys. Res.*, **102**, 9531, 1997.
4. V. Angelopoulos, W. Baumjohann, C. F. Kennel, F. V. Coronti, M. G. Kivelson, R. Pellat, R. J. Walker, H. Luehr, and G. Paschmann. *J. Geophys. Res.*, **97**, 4027–4039, April 1992.
5. W. I. Axford and C. O. Hines: geomagnetic storms. *Can. J. Phys.*, **39**, 1433, 1961.
6. W. B. Baumjohann and R. A. Treumann. *Basic Space Plasma Physics*:
7. L. Biermann: *Z. Astrophys.*, **29**, 274, 1951.
8. J. Birn: *J. Geophys. Res.*, **85**, 1214, 1980.
9. J. Birn, J. F. Drake, M. A. Shay, B. N. Rogers, R. E. Denton, M. Hesse, M. Kuznetsova, Z. W. Ma, A. Bhattacharjee, A. Otto, and P. L. Pritchett: *J. Geophys. Res.*, **106**, 3715, 2001.
10. J. Birn and E. W. Hones: *J. Geophys. Res.*, **86**, 6802–6808, August 1981.
11. J. Birn and K. Schindler: *J. Geophys. Res.*, **88**, 6969–6980, September 1983.
12. J. Birn and K. Schindler: *J. Geophys. Res.*, **107** 18–1, July 2002.
13. J. Birn, K. Schindler, and M. Hesse: *J. Geophys. Res.*, **108**, 2–1, September 2003.
14. J. Birn, R. Sommer, and K. Schindler. *Astrophysics and Space Science*, **35**, 389–402, July 1975.
15. J. Birn, R. R. Sommer, and K. Schindler: *J. Geophys. Res.*, **82**, 147–154, 1977.
16. J. Birn, M. F. Thomsen, J. E. Borovsky, G. D. Reeves, D. J. McComas, R. D. Belian, and M. Hesse: *J. Geophys. Res.*, **102**, 2325–2342, 1997.
17. J. E. Borovsky and H. O. Funsten: the earth's magnetosphere. *J. Geophys. Res.*, **108**, 1246, 2003.
18. W. A. Bristow, A. Otto, and D. Lummerzheim: *J. Geophys. Res.*, **106**, 24593, 2001.
19. W. A. Bristow, G. J. Sofko, H. C. Stenbaek-Nielsen, S. Wei, D. Lummerzheim, and A. Otto: *J. Geophys. Res.*, **108**, 16–1, March 2003.
20. D. Burgess: Collsionless shocks: In M. G. Kivelson and C. T. Russell, editors, *Introductuction to Space Physics*, Cambridge University Press, Ney York, 1995.

21. S. Chandrasekhar: *Hydrodynamic and Hydromagnetic Stability*: Oxford Univ. Press, New York, 1961.
22. S. Chapman and J. Bartels: *Geomagnetism*: Oxford Univ. Press, 1940.
23. S. Chapman and V. C. A. Ferraro: *Nature*, **126**, 129, 1930.
24. Q. Chen, A. Otto, and L. C. Lee: *J. Geophys. Res.*, **102**, 151, 1997.
25. S. H. Chen and M. G. Kivelson: *Geophys. Res. Lett.*, **20**, 2699, 1993.
26. A. J. Dessler: The evolution of arguments regarding the existence of field-aligned currents: In T. A. Potemra, editor, *Magnetospheric currents,* Geophys. Monogr. Ser., vol. 28, 22, AGU, Washington, D. C., 1984.
27. J. W. Dungey: *Phys. Rev. Lett.*, **6**, 47, 1961.
28. A. Egeland: Kristian Birkeland, The man and the scientist: In T. A. Potemra, editor, *Magnetospheric currents,* Geophys. Monogr. Ser., vol. 28, 1, AGU, Washington, D. C., 1984.
29. R. C. Elphic: Observations of flux transfer events, Are fte's flux ropes, islands, or surface waves? In C. T. Russell, E. R. Priest, and L. C. Lee, editors, *Physics of Magnetic Flux Ropes,* Geophys. Monogr. Ser., vol. 58, 455, AGU, Washington, D. C., 1990.
30. R. C. Elphic: Observations of flux tranfer events, A review: In P. Song, B. U. Ö. Sonnerup, and M. F. Thomsen, editors, *Physics of the Magnetopause,* Geophys. Monogr. Ser., vol. 90, 225, AGU, Washington, D. C., 1995.
31. M. Engebretson, N. Lin, W. Baumjohann, , H. Lühr, B. J. Anderson, L. J. Zanetti, T. A. Potemra, R. L. McPherron, and M. G. Kivelson: *J. Geophys. Res.*, **96**, 3441, 1991.
32. G. M. Erickson and R. A. Wolf: *Geophys. Res. Lett.*, **7**, 897–900, November 1980.
33. D. H. Fairfield, A. Otto, T. Mukai, S. Kokubun, R. P. Lepping, J. T. Steinberg, A. J. Lazarus, and T. Yamamoto: *J. Geophys. Res.*, **105**, 21159–21174, 2000.
34. Z. F. Fu, L. C. Lee, and Y. Shi: A three-dimensional mhd simulation of the multiple x line reconnection process: In C. T. Russell, E. R. Priest, and L. C. Lee, editors, *Physics of Magnetic Flux Ropes,* Geophys. Monogr. Ser., vol. 58, 515, AGU, Washington, D. C., 1990.
35. J. T. Gosling, S. J. Bame M. F. Thomsen, R. C. Elphic, and C. T. Russell: *J. Geophys. Res.*, **95**, 8073, 1990.
36. G. Haerendel, G. Paschmann, N. Sckopke, H. Rosenbauer, and P.C. Hedgecock: *J. Geophys. Res.*, **83**, 3195, 1978.
37. E. G. Harris: *Nuovo Cim.*, **23**, 115, 1962.
38. M. Hesse and J. Birn: *J. Geophys. Res.*, **96**, 5683–5696, April 1991.
39. J. R. Kan: *J. Geophys. Res.*, **103**, 11787–11796, June 1998.
40. M. G. Kivelson and C. T. Russell: *Introduction to Space Physics*: Cambridge University Press, New York, 1995.
41. D. Krauss-Varban and N. Omidi: *Geophys. Res. Lett.*, **20**, 1007, 1993.
42. J. La Belle and R. A. Treumann: *Space Sci. Rev.*, **47**, 175, 1988.

43. A. L. La Belle-Hamer, A. Otto, and L. C. Lee: *J. Geophys. Res.*, **100**, 11875, 1995.
44. L. C. Lee: A review of magnetic reconnection, MHD models: In P. Song, B. U. Ö. Sonnerup, and M. F. Thomsen, editors, *Physics of the Magnetopause,* Geophys. Monogr. Ser., vol. 90, 139, AGU, Washington, D. C., 1995.
45. L. C. Lee and Z. F. Fu: *Geophys. Res. Lett.*, **12**, 105, 1985.
46. L. C. Lee and Z. F. Fu: *J. Geophys. Res.*, **91**, 6807, 1986.
47. L. C. Lee, L. Zhang, A. Otto, G. S. Choe, and H. J. Cai: *J. Geophys. Res.*, **103**, 29419–29428, 1998.
48. J. Lemaire: *Planet. Space Sci.*, **25**, 887, 1977.
49. Y. Lin, D. W. Swift, and L. C. Lee: *J. Geophys. Res.*, **101**, 27251, 1996.
50. Z. W. Ma, A. Otto, and L. C. Lee: *J. Geophys. Res.*, **99**, 6125–6136, 1994.
51. M. E. McKean, D. Winske, and S. P. Gary: *J. Geophys. Res.*, **97**, 19,421, 1992.
52. R. L. McPherron: *Reviews of Geophysics and Space Physics*, **17**, 657–681, June 1979.
53. J. E. Midgley and L. Davis: *J. Geophys. Res.*, **68**, 5111, 1963.
54. A. Miura: *Phys. Rev. Lett.*, **49**, 779, 1982.
55. A. Miura: *J. Geophys. Res.*, **89**, 801, 1984.
56. A. Miura and P. L. Pritchett: *J. Geophys. Res.*, **87**, 7431, 1982.
57. P. T. Newell and C.-I. Meng: Magnetosheath injections deep inside the closed llbl, A review of observations: In P. T. Newell and T. Onsager, editors, *Earth's Low-Latitude Boundary Layer,* Geophys. Monogr. Ser., vol. 133, 149, AGU, Washington, D. C., 2003.
58. A. Nishida: *Geophys. Res. Lett.*, **16**, 227, 1989.
59. K. Nykyri and A. Otto: *Geophys. Res. Lett.*, **28**, 3565, 2001.
60. S. Ohtani, K. Takahashi, L. J. Zanetti, T. A. Potemra, R. W. McEntire, and T. Iijima: *J. Geophys. Res.*, **97**, 19311, December 1992.
61. A. Otto: *Comput. Phys. Commun.*, **59**, 185, 1990.
62. A. Otto: *Geophysical and Astrophysical Fluid Dynamics*, **62**, 69, 1991.
63. A. Otto: *J. Geophys. Res.*, **100**, 11863, 1995.
64. A. Otto: *Reviews of Geophysics*, **33**, 657–663, 1995.
65. A. Otto: *Astrophysics and Space Science*, **264**, 17–24, 1999.
66. A. Otto: *J. Geophys. Res.*, **106**, 3751–3758, 2001.
67. A. Otto and D. H. Fairfield: *J. Geophys. Res.*, **105**, 21175, 2000.
68. A. Otto, L. C. Lee, and Z. W. Ma: *J. Geophys. Res.*, **100**, 14895, 1995.
69. A. Otto, K. Schindler, and J. Birn: *J. Geophys. Res.*, 15023–15037, September 1990.
70. E. N. Parker: *J. Geophys. Res.*, **62**, 509, 1957.
71. E. N. Parker: *Astrophys. J.*, **128**, 664, 1958.
72. G. K. Parks: *Physics of Space Plasma*: Addison-Wesley, Reading, MA, 1991.
73. G. Paschmann and P. W. Daly, Eds. *Analysis Methods for Multi-Spacecraft Data*: ESA Publications, Noordwilk, Netherlands, 1998.

74. G. Paschmann, G. Haerendel, I. Papamastorakis, N. Sckopke, S. J. Bame, J. T. Gosling, and C. T. Russell: *J. Geophys. Res.*, **87**, 2159, 1982.
75. G. Paschmann, I. Papamasrtorakis, W. Baumjohann, N. Sckopke, C. W. Carlson, B. U. Ö. Sonnerup, and H. Lühr: *J. Geophys. Res.*, **91**, 11099, 1986.
76. H. G. Petschek: Magnetic annihilation: In W. N. Hess, editor, *AAS-NASA Symposium on the Physics of Solar Flares*. NASA Spec. Publ., SP-50, 425, 1964.
77. E. R. Priest: *Solar Magnetohydrodynamics*: D. Reidel Publ., Dortrecht, Holland, 1987.
78. E. R. Priest and T. Forbes: *Magnetic Reconnection*, MHD Theory and Applications: Cambridge Univ. Press, 2000.
79. T. I. Pulkkinen, D. N. Baker, D. G. Mitchell, R. L. McPherron, C. Y. Huang, and L. A. Frank. *J. Geophys. Res.*, **99**, 5793–5803, April 1994.
80. J. G. Roederer: *Dynamics of Geomagnetically trapped Radiation*: Springer, Berlin, 1970.
81. C. T. Russel and R. C. Elphic: *Space Sci. Rev.*, **22**, 681, 1978.
82. C. T. Russell: The magnetopause: In C. T. Russell, E. R. Priest, and L. C. Lee, editors, *Physics of Magnetic Flux Ropes,* Geophys. Monogr. Ser., vol. 58, 439, AGU, Washington, D. C., 1990.
83. C. T. Russell: A brief history of solar terrestrial physics: In M. G. Kivelson and C. T. Russell, editors, *Introductuction to Space Physics*, 1, Cambridge University Press, Ney York, 1995.
84. C. T. Russell: The structure of the magnetopause: In M. G. Kivelson and C. T. Russell, editors, *Introductuction to Space Physics*, 85, Cambridge University Press, Ney York, 1995.
85. K. Schindler: *Space Sci. Rev.*, **17**, 589–614, June 1975.
86. K. Schindler: *J. Geophys. Res.*, **84**, 7257, December 1979.
87. K. Schindler: *Space Sci. Rev.*, **23**, 365–374, May 1979.
88. K. Schindler: Kinematics of magnetic reconnection in three dimensions: In P. Song, B. U. Ö. Sonnerup, and M. F. Thomsen, editors, *Physics of the Magnetopause,* Geophys. Monogr. Ser., vol. 90, 197, AGU, Washington, D. C., 1995.
89. K. Schindler and J. Birn. *J. Geophys. Res.*, **87**, 2263–2275, April 1982.
90. K. Schindler and J. Birn: *J. Geophys. Res.*, **104**, 25001–25010, November 1999.
91. K. Schindler and J. Birn: *J. Geophys. Res.*, **107**, 20–1, August 2002.
92. K. Schindler and A. Otto: Resistive instability: In C.T. Russell, E.R. Priest, and L.C. Lee, editors, *Physics of Magnetic Flux Ropes,* Geophys. Monogr. Ser., vol. 58, 51–61, AGU, Washington, D. C., 1990.
93. M. Scholer: *Geophys. Res. Lett.*, **15**, 291, 1988.
94. M. Scholer, M. Fujimoto, and H. Kucharek: *J. Geophys. Res.*, **98**, 18971, 1993.

95. N. Sckopke, G. Paschmann, G. Haerendel, B. U. Ö. Sonnerup, S. J. Bame, T. G. Forbes, E. W. Hones, Jr., and C. T. Russell: *J. Geophys. Res.*, **86**, 2099, 1981.
96. V. A. Sergeev, R. J. Penninen, and T. I. Pulkkinen: *Space Sci. Rev.*, **75**, 551–604, February 1996.
97. V. A. Sergeev, P. Tanskanen, K. Mursula, A. Korth, and R. C. Elphic: *J. Geophys. Res.*, **95**, 3819–3828, April 1990.
98. M. A. Shay, J. F. Drake, B. N. Rogers, and R. E. Denton: *J. Geophys. Res.*, **106**, 3759, 2001.
99. D. G. Sibeck: *J. Geophys. Res.*, **94**, 2543, 1990.
100. D. G. Sibeck: *J. Geophys. Res.*, **97**, 4009, 1992.
101. B. U. Ö. Sonnerup: *J. Geophys. Res.*, **85**, 2017, 1980.
102. B. U. Ö. Sonnerup and L. J. Cahill, Jr.: *J. Geophys. Res.*, **72**, 171, 1967.
103. B. U. Ö. Sonnerup, I. Papamastorakis, G. Paschmann, and H. Lühr: *J. Geophys. Res.*, **92**, 12137, 1987.
104. B. U. Ö. Sonnerup, I. Papamastorakis, G. Paschmann, and H. Lühr: *J. Geophys. Res.*, **95**, 10541, 1990.
105. B. U. Ö. Sonnerup, G. Paschmann, I. Papamastorakis, N. Sckopke, G. Haerendel, S. J. Bame, J. R. Asbridge, J. T. Gosling, and C. T. Russel: *J. Geophys. Res.*, **86**, 10049, 1981.
106. D. J. Southwood, C. J. Farrugia, and M. A. Saunders: *Planet. Space Sci.*, **36**, 503, 1988.
107. R. G. Stone and B. T. Tsurutani: *Collisionless Shocks in the Heliosphere*, A tutorial review, Geophys. Monogr. Ser., vol. 34, AGU, Washington, D. C., 1985.
108. P. A. Sweet: The neutral point theory of solar flares: In B. Lehnert, editor, *Electromagnetic Phenomena in Cosmical Physics*. Cambridge University Press, 1958.
109. R. A. Treumann and W. B. Baumjohann: *Advanced Space Plasma Physics*: Imperial College Press, London, 1997.
110. B. T. Tsurutani, W. D. Gonzalez, Y. Kamide, and J. K. Arballo: *Magnetic Storms,* Geophys. Monogr. Ser., vol. 98: AGU, Washington, D. C., 1997.
111. N. A. Tsyganenko: *Space Sci. Rev.*, **54**, 75–186, 1990.
112. N. A. Tsyganenko: *J. Geophys. Res.*, **105**, 27739–27754, December 2000.
113. N. A. Tsyganenko and D. P. Stern: *J. Geophys. Res.*, **101**, 27187–27198, December 1996.
114. V. M. Vasyliunas: *Space Sci. Rev.*, **13**, 303, 1975.
115. R. J. Walker and C. T. Russell: Solarwind interactions with magnetized planets: In M. G. Kivelson and C. T. Russell, editors, *Introductuction to Space Physics*, 1, Cambridge University Press, Ney York, 1995.
116. R. A. Wolf: Magnetospheric configuration: In M. G. Kivelson and C. T. Russell, editors, *Introductuction to Space Physics*, 288, Cambridge University Press, Ney York, 1995.
117. B. J. Zwan and R. A. Wolf: *J. Geophys. Res.*, **81**, 1636, 1976.

Space Weather Effects in the Upper Atmosphere: Low and Middle Latitudes

Gerd W. Prölss

Institut für Astrophysik und Extraterrestrische Forschung, Universität Bonn, 53121 Bonn, Germany

Abstract. In this chapter the space weather effects in the upper atmosphere are investigated. It is discussed how space weather affects satellites and radio transmissions. Emphasis is put on severe weather conditions during so-called upper atmospheric storms.

1 Introduction

It is well known that the conditions in the space environment of the Earth are constantly changing. This variability has become known as 'space weather'. It is implied that conditions in space are as difficult to predict as tropospheric weather. This term is also consistent with the well-established practice of designating large perturbations of the space environment as 'storms'. The first to use such terminology was Alexander von Humboldt. When describing large and long-lasting disturbances of the Earth's magnetic field he used the term 'Magnetisches Ungewitter' (Humboldt, 1808 [30]; Humboldt, 1845 [31]). 'Ungewitter' means 'bad weather' or 'storm', and this latter term was used by E. Sabine when he translated Humboldt's monumental scientific work *Kosmos* into English. Today the term 'Magnetischer Sturm' is also used in German.

It is difficult to establish who first used the term 'space weather'. The present author first came across this term in a report by T.M. George on ionospheric effects of atmospheric waves published in 1967. It should also be noted that today we are not only confronted with 'space weather' but also with 'galactic weather' with cold and warm fronts (gas clouds) interacting with each other.

In what follows we will investigate the space weather in the upper atmosphere and how it affects satellites and radio transmissions. By focusing on severe weather conditions, i.e. on 'upper atmospheric storms', their effects and their underlying physics become more clear. We will introduce the subject matter by briefly reviewing some basic properties of the upper atmosphere.

2 Thermospheric Storms

The upper atmosphere comprises the outer region of the terrestrial gas envelope above about 100 km altitude. As indicated in Fig. 1 this region exhibits

G.W. Prölss, Space Weather Effects in the Upper Atmosphere: Low and Middle Latitudes, Lect. Notes Phys. **656**, 193–214 (2005)
http://www.springerlink.com/

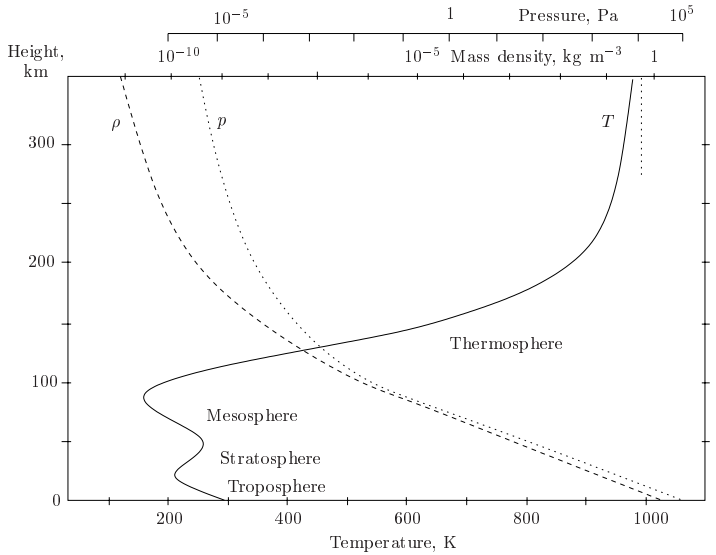


Fig. 1. Representative height profiles of the temperature (T), pressure (p), and mass density (ρ) in the Earth's atmosphere.

a large increase in the gas temperature to approximately 1000 K. Accordingly, this region is also called 'thermosphere'. The increase in temperature is caused by the absorption of ultraviolet radiation from the sun. Fortunately, an astronaut working in this environment will not suffer from this heat. This is because the density at these altitudes has decreased so much that one is effectively working under high vacuum conditions. As indicated in Fig. 1 the mass density has decreased by more than ten orders of magnitude at 300 km altitude. Nevertheless, there are still some 10^{15} gas particles per cubic meter left at this height. This is enough to exert a significant drag force on any object moving through this region with high velocity like, for example, a satellite in low earth orbit or the International Space Station. In fact this drag is so strong that these objects constantly lose height, up to several hundred meters per day.

Solar radiation is not the only heat source of the upper atmosphere. As discussed in previous contributions and as indicated schematically in Fig. 2, solar wind kinetic energy is partly captured by the Earth's magnetosphere via a magnetoplasmadynamic generator process. This way solar wind kinetic energy is transformed into electromagnetic energy and subsequently transferred to the polar region by electric currents and accelerated particles. A very important property of this solar wind energy source is its high variability. There are times when relatively little energy is dissipated in the polar regions. Such times are called 'quiet'. And there are 'active' or 'disturbed' times when large amounts of energy are deposited in these regions. Since it

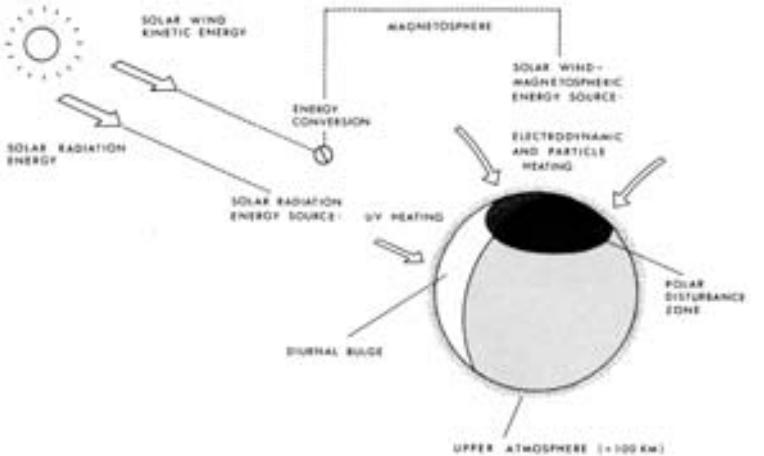


Fig. 2. Principle modes of solar energy transfer to the terrestrial upper atmosphere. Note that electrodynamic (Joule) heating dominates the energy dissipation at polar latitudes.

is quite difficult to measure the rate of this energy deposition directly, it is common practice to use magnetic activity indices like the Dst or AE indices to monitor the actual strength of the solar wind energy source.

The strong heating during disturbed conditions not only affects the high- but also the mid- and low-latitude upper atmosphere. This will be documented using density data obtained by the CASTOR satellite during the storm event of January 10/11, 1976. The CASTOR satellite was in a low inclination orbit with a perigee height of about 275 km (e.g. Villain, 1980 [70]; Berger and Barlier, 1981 [6]). During the storm event considered, mass density data were collected near 25 °N in the local midnight sector. In order to keep the measurement errors at an acceptable level, only density data obtained below 400 km altitude were considered. These data were adjusted to a common altitude of 325 km using standard aerostatic technics and were subsequently averaged. This way one mean density value for an altitude of 325 km altitude was obtained for each perigee pass.

Using the Dst index, the upper part of Fig. 3 describes the magnetic activity during the January storm. The steep drop of this index between noon and midnight on January 10 testifies to the large amount of energy deposited in the inner magnetosphere during this time interval. Following this energy injection, a large increase in the upper atmospheric gas density was observed at low latitudes; see lower part of Fig. 3. As compared to prestorm conditions, the mass density increased by nearly 300%! This will, of course, increase the drag on a satellite or the International Space Station passing through this high density region. Accordingly, their orbital elements will suddenly change, rendering any ephemeris predictions obsolete. It is also easy to visualize that

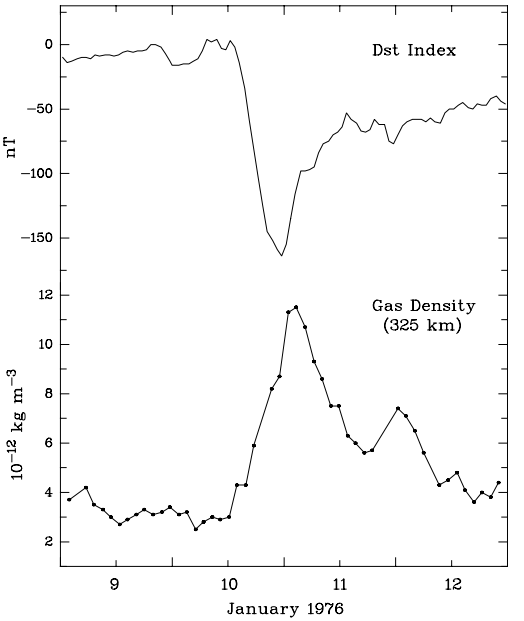


Fig. 3. Magnetic storm associated density perturbations at low latitudes. Using the Dst index the upper part illustrates the development of magnetic activity during the storm event of January 10/11, 1976. The associated perturbation in the mass density of the upper atmosphere at an altitude of 325 km is shown in the lower part of the figure. The data refer to a mean latitude of 25 °N and were obtained in the local midnight sector.

if such storms would occur more frequently (like during high solar activity conditions), they would considerably shorten the lifetime of a spacecraft. In particularly severe cases, the increased drag may even cause the total loss of a satellite. This happened, for example, to the Advanced Satellite for Cosmology and Astrophysics (ASCA), an X-ray astronomy satellite, during the large storm of July 2000; see the e-mail reprinted in an abbreviated form below.

September 20, 2000

Dear ASCA users,
As was announced in July, ASCA has transferred into a safe-hold mode (SH-mode). In spite of our serious efforts for about two months, we are not able to recover ASCA to the normal observation mode. Here, we wish to report briefly on the disturbance event and current status of ASCA. A historically big solar flare occurred on July 14, 2000, with extremely strong solar proton flux. Subsequently, the disturbance triggered a big geomagnetic storm that continued for a few days after July 15. Due to the geomagnetic storm, the atmosphere sporadically expanded and the atmospheric gas density at the ASCA altitude suddenly increased to several times the normal value. This

caused the external torque on the satellite due to the air drag to increase. Thus, the ASCA attitude was perturbed and the on-board attitude control system (ACS) transferred ASCA to the SH-mode. Since the accumulated external torque was stronger than the compensable internal torque stored in the on-board momentum wheels, the ASCA attitude was not locked to the nominal aspect of SH-mode, and further moved away the solar paddles from the nominal direction normal to the sun. This reduced power generation by the cells and finally exhausted the battery power completely. After the event, we attempted all possible and considerable operations to recover the aspect and to charge up the battery. However, there has been no improvement so far. We suspect that the battery cells may have suffered serious unrecoverable damage. Taking all of the situation explained above into account, we regret that we must announce that the possibility that ASCA will return to observation mode is very small, almost hopeless, even though we will monitor the ASCA status until its re-entry into the atmosphere. We appreciate your understanding that we must cancel all the programmed long observations after July 15, 2000.

H.Inoue, F.Nagase and ASCA operation team

How do we explain such density perturbations? Usually, they are attributed to a heating and subsequent expansion of the upper atmosphere, as indicated in the ASCA report. If this explanation is accepted, we are then faced with the question of what causes this heating. As to the low latitude region considered in Fig. 3, this question is not readily answered and several mechanisms have been proposed; see Table 1. Today, the first three of the *local* heating mechanisms listed in this table are only of historical interest, the fourth is still of general interest. The disturbance scenario envisaged in this latter case is sketched in Fig. 4.

It is well known that a large number of energetic particles (energy range 1 to 200 keV) is injected into the magnetospheric ring current during magnetic storms. There they form a considerable reservoir of energy ($\simeq 10^{16}$ J). Part of this energy will be transmitted to the upper atmosphere by energetic neutral atoms (ENA's) formed in charge-exchange collisions with ambient exospheric hydrogen particles. It is the energy dissipation by these precipitating particles which has been made responsible for the heating of the upper atmosphere at middle and low latitudes.

In the meantime it was demonstrated that this mechanism does explain some unusual airglow emissions (e.g. Stephan et al., 2000 [62]) and a special kind of ionospheric disturbance effect (Bauske et al., 1997 [4]) but it is not sufficient to explain the heating effects (Prölss et al., 1973 [56]; Noël and Prölss, 1993 [44]). This then leaves us with the *non-local* heating mechanisms. Here it is assumed that the energy is first deposited at high latitudes and subsequently transported towards lower latitudes. Whether this energy transport is primarily affected by large-scale convection or by waves is still an open question. However, it is reasonable to assume, that at least during the

initial phase of a storm, the energy transported by waves is more important. The disturbance scenario envisaged in this case is outlined in Fig. 5.

During a magnetospheric substorm (indicated here by an increase in the AE index) a considerable amount of energy is injected into the polar upper atmosphere. This sudden heat addition generates a whole spectrum of atmospheric gravity wavesndexwave!gravity. At middle latitudes these waves superimpose to form an impulse-like travelling atmospheric disturbance (TAD) which propagates with high velocity towards lower latitudes. At the equator the TADs originating in the southern and northern hemispheres clash and cause a compression and heating of the atmospheric gases; see lower part of Fig. 5.

There is some evidence in support of this disturbance scenario. First we note that the density perturbations at low latitudes are at least partially caused by a compression of the gases. This follows from the behavior of helium whose density may increase by more than 50%. For such a light gas constituent this increase is much too large to be solely explained as a temperature effect, especially since the density enhancement of the heavier gas constituent molecular nitrogen indicates a rather modest temperature rise (Prölss, 1982 [50]). There is also some evidence supporting the existence of the equatorward-directed winds held responsible for the compression of the thermospheric gases. At middle latitudes such winds should produce positive ionospheric storms. In order to better understand this kind of disturbance effect let us first recall some basic properties of the ionosphere.

Table 1. Heating mechanisms for the low-latitude upper atmosphere

1. Local heating of the equatorial upper atmosphere by
   (a) flash of UV radiation (Maris and Hulburt, 1929 [38])
   (b) dissipation of MHD and heat conduction waves (Dessler, 1959 [17]; Volland, 1967 [71])
   (c) joule dissipation of ionospheric currents (Cole, 1962 [13])
   (d) precipitation of neutralized ring current particles (e.g. Evans, 1970 [20]; Tinsley, 1981 [69]; Bencze et al., 1992 [5])
2. Heating of the polar upper atmosphere and subsequent transport of energy towards lower latitudes by
   (a) large-scale wind circulation (e.g. Johnson, 1960 [33]; Volland and Mayr, 1971 [72]; Richmond, 1979 [58])
   (b) waves/travelling atmospheric disturbances (TADs) (e.g. Gold, 1963 [28]; Blumen and Hendl, 1969 [7]; Testud, 1970 [66]; Klostermeyer, 1973 [35]; Cole and Hickey, 1981 [14]; Prölss, 1982 [50]; Burns and Killeen, 1992 [9]; Prölss, 1993 [51]; Fujiwara et al., 1996 [24])

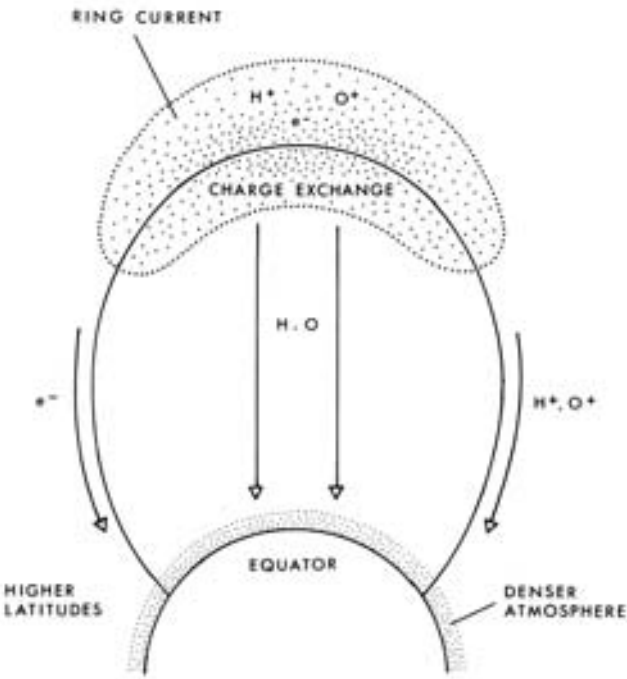


Fig. 4. Heating of the low latitude upper atmosphere by precipitating neutralized ring current particles. Also indicated is the precipitation of energetic electrons from the outer radiation belt along the magnetic field lines at higher mid-latitudes.

3 Ionospheric Storms

The ionosphere comprises the ionized component of the upper atmosphere. It is formed by the ionizing action of the sun's radiation in the extreme ultraviolet and X-ray range. As indicated in Fig. 6, the ionization is mainly concentrated in a layer whose thickness is of the order of a few hundred kilometers and whose peak is located near 300 km altitude. Here atomic oxygen ions and electrons are the major constituents. The maximum ionization density is of the order of 10^{12} m^{-3}, and this constitutes the largest concentration of charged particles in the Earth's space environment. Nevertheless, even at the layer peak, neutral gas particles outnumber ions and electrons by 1000 to 1. Evidently, we are dealing with a weakly ionized gas.

In spite of being a trace constituent, the ionosphere completely changes the electromagnetic properties of the upper atmosphere. For example, radio waves passing through this region are now refracted or even reflected. Reflection occurs wherever the radio frequency reaches the local resonance frequency of the ionosphere, i.e. the plasma frequency, f_P [Hz] $\simeq 9(n\ [\mathrm{m}^{-3}])^{1/2}$.

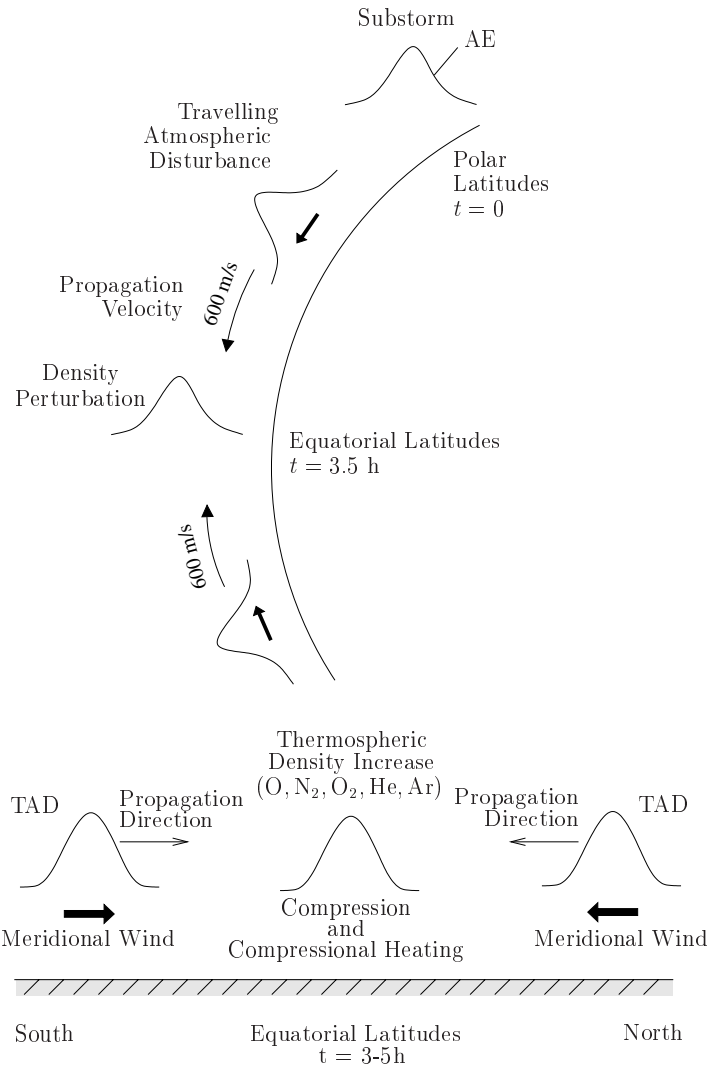
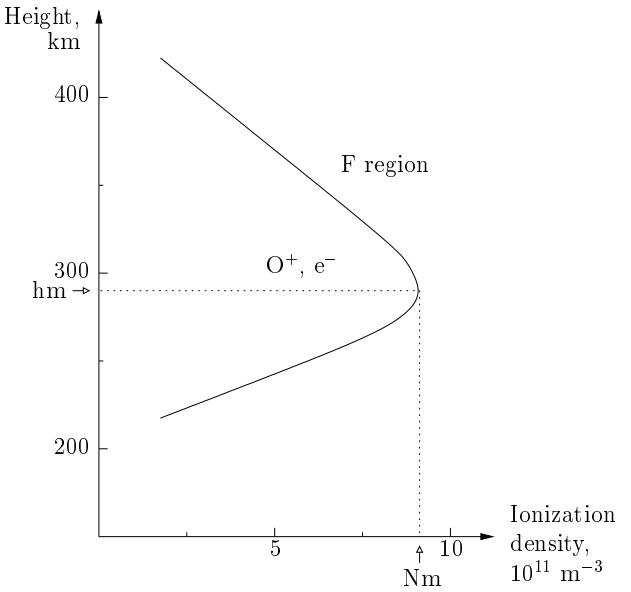


Fig. 5. Causal relationship between magnetospheric substorms (AE index), travelling atmospheric disturbances (TADs), and density perturbations at low latitudes. The explicit values for the propagation velocity and arrival times serve only to illustrate the order of magnitude of these quantities. Also, the time variation of a TAD is more complicated than is indicated in this scheme.

Here we note that it was the reflection of radio waves which first established the existence of the ionosphere. Also, it was this reflection of radio waves which formed the basis of long-distance communication during a large part of the 20th century. Finally, it was this kind of radio communication which experienced one of the first space weather effects.



Ionization production: $O + \text{solar radiation } (\lambda \leq 100 \text{ nm}) \longrightarrow O^+ + e^-$

Ionization loss:

$$O^+ + N_2 \longrightarrow NO^+ + N$$
$$O^+ + O_2 \longrightarrow O_2^+ + O$$
$$(\text{subsequently: } NO^+,\ O_2^+ + e^- \longrightarrow N, O; \quad \text{fast!})$$

Fig. 6. Representative ionization density profile at middle latitudes, and major ionization production and loss processes in the F region.

It all started in the early 1920s when experiments in one-way radio telephone transmissions were conducted by the Bell Telephone Company. As indicated in Fig. 7, radio signals were transmitted from Rocky Point near New York to New Southgate near London over a total distance of 5482 km. The transmission frequency used was 57 kHz, which corresponds to a wave length of about 5 km. During these experiments a definite correlation was found between abnormal radio transmissions and disturbances of the Earth's magnetic field (Espenschiedt et al., 1925 [19]). The main effect observed was a significant decrease of the nighttime field strength; see Fig. 7. Subsequent studies confirmed this disturbance effect and even tried to directly correlate it with solar activity (e.g. Pickard, 1927a [47]; Pickard, 1927b [48]; Pickard, 1927c [49]; Anderson, 1928 [1]; Anderson, 1929 [2]; Wymore, 1929 [73]). Incidently, none of the studies cited even mentions the ionosphere or the *Kennelly-Heaviside layer* as it was known at that time. This is because in

the 1920s and especially among radio engineers this layer was still considered an academic myth rather than reality.

Today we know that during and especially following magnetic storms, energetic electrons from the outer radiation belt precipitate along magnetic field lines into the mid-latitude upper atmosphere; see Fig. 4. There they produce extra ionization in the lower ionosphere (D-region) which in turn enhances the attenuation of radio waves passing through this region (e.g. Lauter, 1961 [36]; Bremer, 1998 [8]). Evidently, this mechanism does explain the nighttime observations documented in Fig. 7.

Nowadays, this kind of storm effect is only of secondary importance. This is because generally much higher frequencies are used which are less sensitive to such attenuation effects, and because most of the time the reflecting ionosphere is replaced by communication satellites in geostationary orbit. This does not mean, however, that the ionosphere has become unimportant for the transmission of radio waves. On the contrary, the more sophisticated such transmission systems become, the more they depend on ionospheric conditions. A good example is the global navigation satellite systems (GNSS). While passing through the dispersive ionosphere, the GNSS signals are slightly delayed. For one-frequency transmissions, this results in ranging errors of about 50 m which need to be corrected. The actual magnitude of this error will depend on the total electron content along the signal path. Now the total electron content of the ionosphere is highly variable, especially during disturbed conditions. This is illustrated in Fig. 8 for the moderately strong storm of May 15-1972. Again magnetic activity indices are used to describe the level of solar wind energy deposition. In response to the enhanced energy addition on May 15, significant increases in both the total electron content (TEC) and the maximum ionization density (Nm) are observed. These so-called *positive ionospheric storm* effects are followed by a depression of the ionization density (i.e. *negative ionospheric storm* effects) on May 16. It is easy to visualize that such unexpected changes will adversely affect the accuracy of the GNSS ranging system. To make things worse, storms frequently cause rapid fluctuations of the amplitude and phase of the GNSS signals (scintillations) which may prevent a position-fixing altogether.

How do we explain such perturbations? In the case of positive ionospheric storms we turn again to travelling atmospheric disturbances. As noted previously, such disturbances are associated with equatorward-directed winds of moderate magnitude. In Fig. 9, for example, their amplitude is assumed to be of the order of 150 m/s; this value should not be confused with the much larger propagation velocity ($\simeq$ 600 m/s). Here we are interested in the frictional force the meridional winds exert on the ionization. Since charged particles are free to move only along the magnetic field, the wind will push the ionization upward along the inclined field lines of the Earth's magnetic field. This way the ionosphere is moved to higher altitudes. Now the ionization losses – which are proportional to the molecular nitrogen and oxygen densities in the upper atmosphere – decrease much faster with altitude than

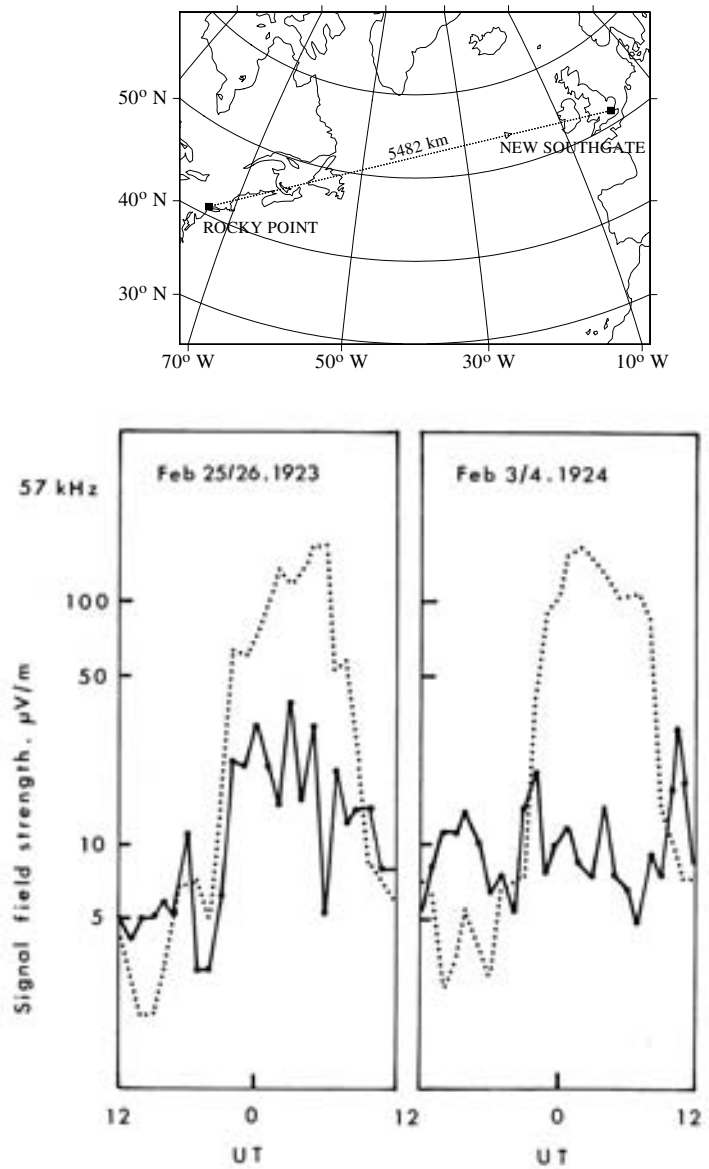


Fig. 7. Transmission path of early transatlantic radio telephone experiments (upper panel) and magnetic storm associated perturbations of the received signal field strength (lower panels). The signal field strength is plotted as a function of Universal Time. The dotted lines refer to the undisturbed conditions on February 22/23-1923 (left hand panel) and January 27/28-1924 (right hand panel), respectively. The continuous lines show the storm time variations on the days indicated. (Lower panels adapted from Espenschiedt et al., 1925 [19]).

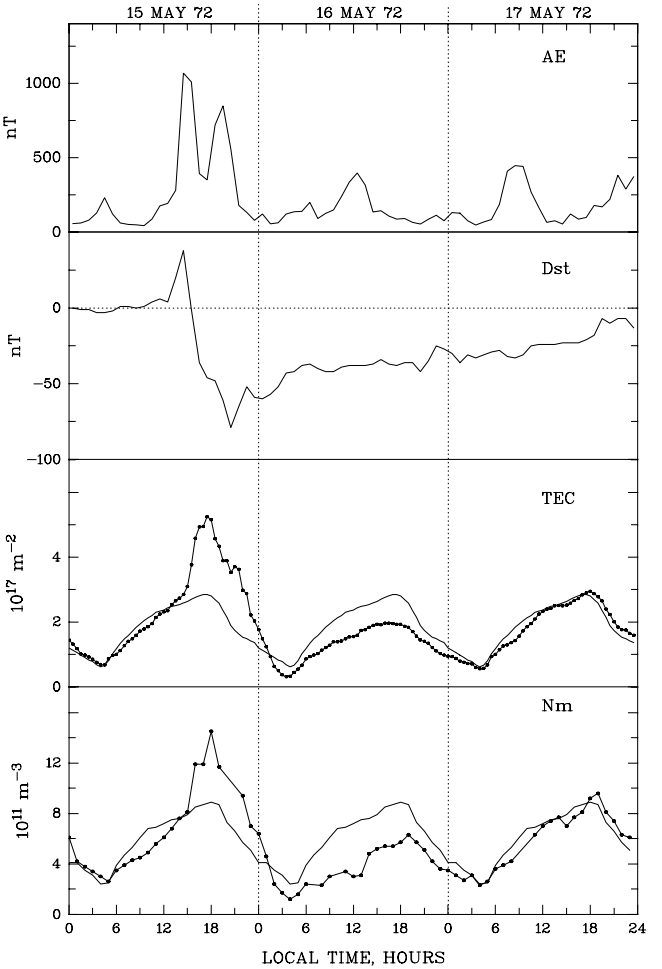


Fig. 8. Ionospheric response to a magnetic storm in May 1972. The AE and Dst indices are used to indicate the level of magnetic activity during this event. The two lower panels show the associated changes in the total electron content (TEC) and in the maximum electron density (Nm) of the ionosphere, respectively. The curves labeled with small solid circles indicate the storm time variations whereas the thin solid curves serve as a quiet time reference. This reference corresponds to the mean variation observed on the seven days prior to the storm event. The TEC data were obtained at Hamilton using the Faraday polarization twist of VHF radio waves transmitted from the geostationary satellite ATS-3. At an altitude of 420 km the subionospheric coordinates are 38.9 °N, 70.7 °W. The maximum ionization density measurements were recorded at the ionosonde station Wallops Island (37.9 °N, 75.5 °W). The ionospheric data shown in this figure are taken from Mendillo and Klobuchar (1974) [39].

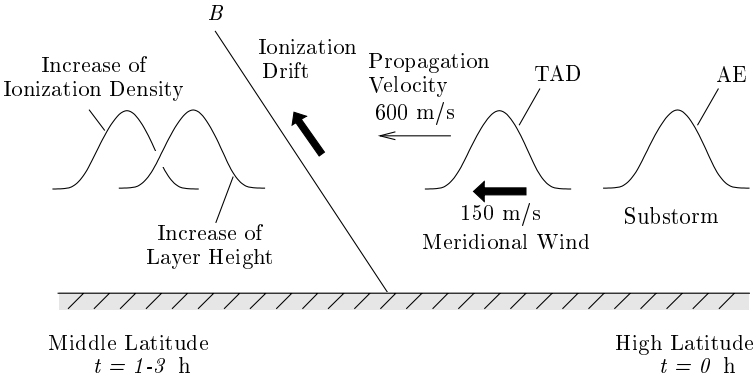


Fig. 9. Causal relationship between magnetospheric substorms (AE index), traveling atmospheric disturbances (TAD) and short-duration positive ionospheric storms (increase of layer height and ionization density). The explicit values for the propagation velocity and the meridional wind velocity of the TAD and the arrival times serve only to illustrate the order of magnitude of these quantities. Also, the time variation of a TAD is more complicated than is indicated in this scheme (Prölss and Ocko, 2000 [54]).

the ionization production which is proportional to the atomic oxygen density; see Fig. 6. Therefore the uplifting of the ionosphere will cause, with a certain time delay, an increase in the ionization density, i.e. a positive ionospheric storm.

The disturbance scenarios described in Figs. 5 and 9 imply that whenever one observes larger positive ionospheric storms at middle latitudes, one should also observe neutral density perturbations at equatorial latitudes, and vice versa. As it turns out, data sets which could verify such a correlation are scarce and Fig. 10 presents one of the few examples available (for further examples see, Prölss, 1993 [51]). In response to the strong substorm activity observed on April 1-1976, and with a certain time delay, first an increase in the layer height and then an increase in the ionization density (ΔNm) is recorded at middle latitudes. After a further time delay, an increase in the thermospheric gas density is observed at low latitudes.

Winds are not the only way to generate positive ionospheric storms. Other mechanisms include neutral composition changes (e.g. Danilov et al., 1987 [16]; Burns et al., 1995 [10]; Mikhailov et al., 1995 [42]; Fuller-Rowell et al., 1996 [25]; Field et al., 1998 [21]; Strickland et al., 2001b [64]), horizontal ionization transport (Foster, 1993 [23]), and electric fields (e.g. Tanaka, 1986 [65]; Jakowski et al., 1992 [32]; Pi et al., 1993 [46]). In particular, the significance of electric fields for generating storm effects at middle latitudes and also their origin are still hotly debated topics in the literature (see also Scherliess and Fejer, 1998 [60]).

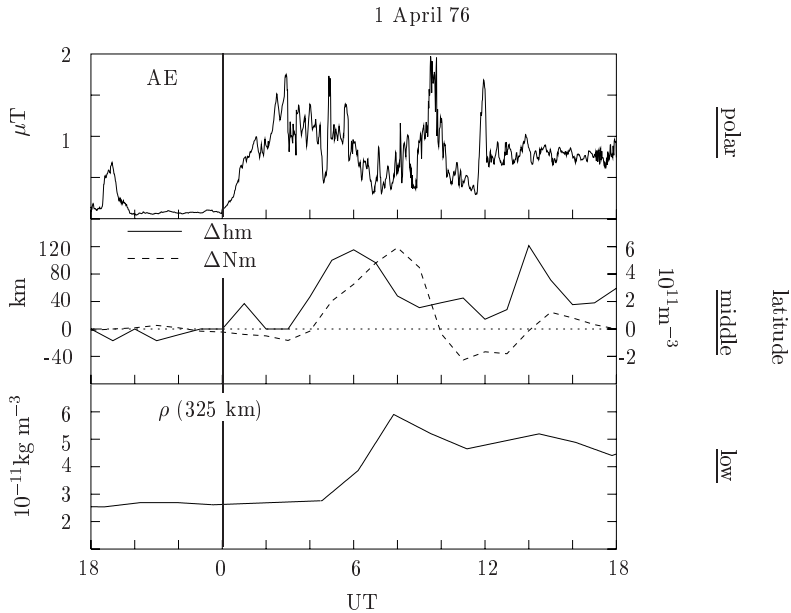


Fig. 10. The April 1-1976 storm event. Shown from top to bottom are: (top) substorm activity, as indicated by the AE(10) index; (middle) storm-induced changes in the height of the peak electron density of the F2 layer, Δhm, and in the associated peak electron density, ΔNm, as observed at the mid-latitude station Alma Ata (43 °N, 77 °E); (bottom) geomagnetic activity effect on the thermospheric mass density, ρ, as observed by the CASTOR satellite at low latitudes. Note that the latter density values refer to an altitude of 325 km, an approximate latitude of 10 °N, and to the early afternoon local time sector. (Prölss and Ocko, 2000 [54])

Much less controversial is the origin of negative ionospheric storms. Here observations clearly demonstrate that changes in the neutral gas composition are responsible for such disturbance effects (e.g. Prölss, 1995 [52]). Figure 11 illustrates this causal relationship using data obtained during the severe storm of July 14-1982. The upper panel of this figure documents the relative changes in the molecular nitrogen density and in the atomic oxygen density observed during this event. In this kind of presentation $R(n) = 1$ means no change with respect to reference conditions. A prominent feature of this data set is the neutral composition disturbance zone which extends all the way from high to low latitudes. It is marked by a significant increase in the molecular nitrogen density and a concurrent decrease in the atomic oxygen density.

Both density perturbations have important implications for the ionosphere. This is because a decrease in the atomic oxygen density will decrease the production of ionization and an increase in the molecular nitrogen density will increase the loss of ionization; see Fig. 6. Thus both changes combine to reduce the ionization density. Any ionosonde station located below such a

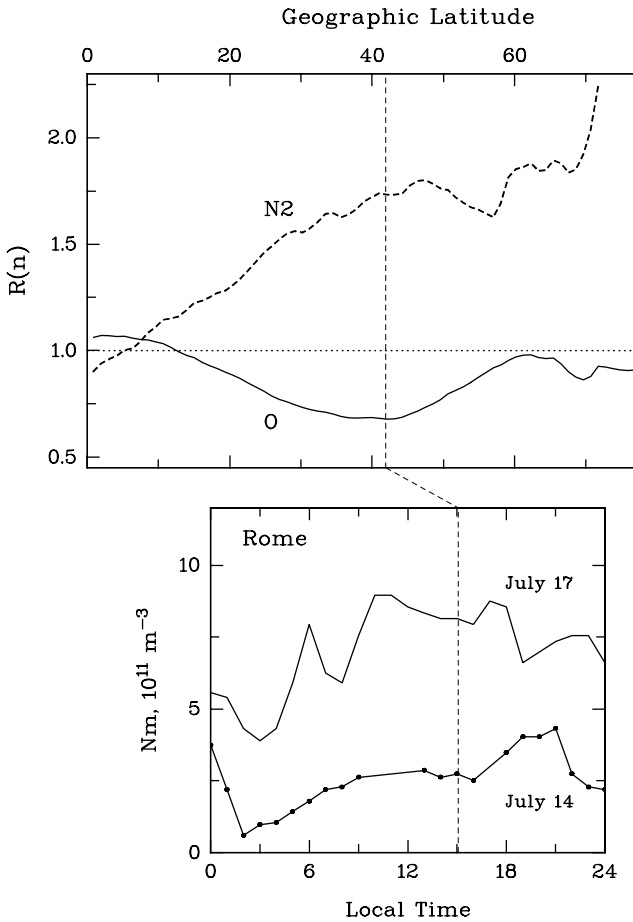


Fig. 11. Correlation between neutral composition changes and negative ionospheric storms. The upper panel presents storm induced perturbations in the molecular nitrogen (N_2) and in the atomic oxygen (O) densities as measured by the DE-2 satellite on July 14-1982. Relative changes are plotted, and R(n)=1 serves as a reference meaning no changes with respect to the reference conditions observed on July 17-1982. The data are plotted as a function of geographic latitude and the location of the ionosonde station Rome is indicated by a vertical broken line. At this latitude the satellite data were obtained near 15 hours solar local time, 18 °E geographic longitude, and at about 320 km altitude. The lower part of the figure shows the local time variation of the maximum electron density of the ionosphere as observed at the ionosonde station Rome (41.9 °N, 12.5 °E). Storm-time data (July 14-1982) are compared to reference data (July 17-1982). The time of the satellite measurements are indicated by a vertical broken line. For further details see Prölss and Werner (2002) [55].

composition disturbance zone should therefore record a decrease in the ionization density. The lower part of Fig. 11 confirms this supposition. Plotted is the local time variation of the maximum electron density as observed at Rome. Measurements obtained on the storm day July 14 are compared to those recorded on the reference day July 17. As is evident, large negative storm effects are observed on the storm day, in agreement with expectations.

A simple estimate indicates that the relative change in the ionization density should be of the same order of magnitude as the relative change in the O/N_2 density ratio, $R(N_m) \simeq R(O/N_2)$. For the specific case illustrated in Fig. 11 we have $R(O/N_2) \simeq 0.39$, and this is reasonably close to the value of $R(N_m) \simeq 0.33$. More accurate results are obtained if the observed composition changes are used as an input to a numerical simulation of the ionospheric behavior.

In situ measurements of changes in the thermospheric gas densities are restricted to the location of the satellite trajectory. Supplementary information on the *global* distribution of density perturbations is provided by remote satellite observations of the far-ultraviolet (FUV) dayglow from high above the upper atmosphere. A significant fraction of this dayglow consists of scattered solar radiation, which is similar to what we learned from the lower atmosphere. In the FUV range this radiation is dominated by emissions from atomic oxygen at about 130.4 nm (more precisely by the triplet at 130.2, 130.5, and 130.6 nm). These prominent emission lines are also excited by inelastic collisions with photoelectrons. In any case, the emission intensity will depend on the atomic oxygen density in the upper atmosphere. Here we compare the spatial extent of oxygen depletion areas inferred from dayglow images with concurrently observed distributions of negative ionospheric storm effects.

The central part of Fig. 12 documents changes in the FUV dayglow intensity observed during the storm event of September 27-1981. Dayglow changes are shown in the form of percent differences with the color bar at the bottom of the image indicating the magnitude of these deviations. Note that the unusual shape of this image results from the observing geometry and from the exclusion of the aurora. Also indicated are the locations of the four ionosonde stations Lannion, Sverdlovsk, Tomsk, and Khabarovsk. As can be seen, significant decreases in the airglow intensity are observed above Eurasia during the storm event.

The local time variations of the maximum electron density (NmF2) as observed at the four ionosonde stations are shown in the four panels surrounding the satellite image. Filled and open circles refer to quiet time and storm time conditions, respectively. The time at which the satellite image was taken is indicated by a vertical broken line.

A comparison of the two data sets shows that the two stations Lannion and Khabarovsk, located below the region of depressed airglow, observe a reduction of the ionization density, whereas Sverdlovsk and Tomsk, located outside this region, record normal variations. Since a depressed airglow intensity

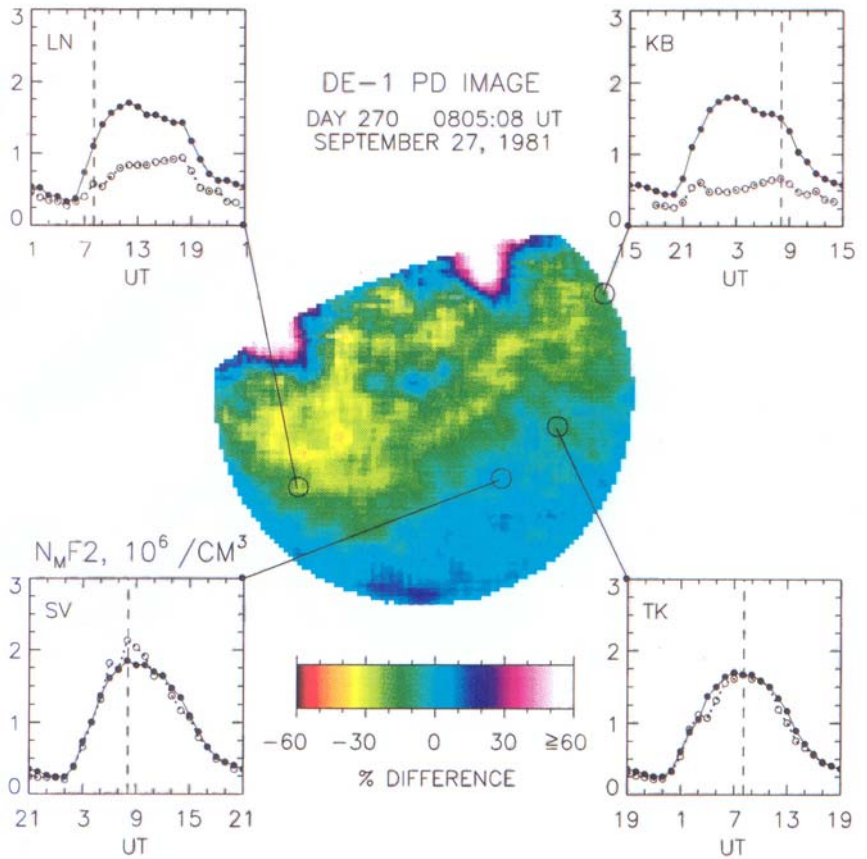


Fig. 12. Spatial correlation between depressions of the FUV dayglow and negative ionospheric storm effects. The central image of this figure shows the percentage decrease in the FUV dayglow (123-160 nm) as observed by the DE-1 satellite on September 27-1981, at about 0805 UT. The central meridian longitude is 60 °E at ∼ 12 LST. Dayglow changes are shown in the form of percent differences, where the color bar at the bottom of the figure indicates the magnitude of these deviations. Also indicated are the sites of four Eurasian ionosonde stations: Lannion (LN; 49 °N, 357 °E), Sverdlovsk (SV; 56 °N, 59 °E), Tomsk (TK; 57 °N, 85 °E), and Khabarovsk (KB; 49 °N, 139 °E). The local time variation of the maximum electron density of the F2 layer (NmF2) as observed at these stations on September 27-1981, is shown in the four panels surrounding the satellite image (open circles). The units are 10^6 electrons/cm^3. A non-storm reference, the 30-day median of NmF2 centered on September 27, is also shown (filled circles). The time at which the satellite image was taken is indicated by the vertical dashed lines. The observation height of the DE-1 satellite was about 21 000 km Prölss and Craven, 1998 [53].

corresponds to a reduced atomic oxygen density, this comparison confirms the close correlation between thermospheric density changes and negative ionospheric storm effects, this time on a global scale.

At first sight it may come as a surprise that moderately strong depressions of the dayglow are associated with fairly large negative ionospheric storm effects. Here we note that any storm-induced decrease in the atomic oxygen density is invariably associated with a concurrent increase in the molecular nitrogen (and molecular oxygen) density, which further decreases the ionization density; see Fig. 11. Also, the decrease in the oxygen density is somewhat larger than indicated by the dayglow changes (e.g. Strickland et al., 2001a [63], Strickland et al., 2001b [64]). Quantitative results are obtained by numerically simulating the disturbance effects.

4 Simulation of Upper Atmospheric Storms

Numerous models of varying degrees of complexity and sophistication have been developed for the simulation of upper atmospheric storm effects. Among them are models which focus on the disturbed thermosphere, others which deal with the disturbed ionosphere, and those which attempt to reproduce both the thermospheric and ionospheric storm behavior. Thermospheric models are used, for example, to investigate the properties of travelling atmospheric disturbances. The basic equation to be solved in this case is the equation of motion

$$\rho \frac{\partial \boldsymbol{u}_n}{\partial t} + \rho_n (\boldsymbol{u}_n \boldsymbol{\nabla}) \boldsymbol{u}_n = - \boldsymbol{\nabla} p_n + \eta (\Delta \boldsymbol{u}_n + \boldsymbol{\nabla} (\boldsymbol{\nabla} \boldsymbol{u}_n)/3) + \rho_n \boldsymbol{g} + \rho_n \nu^*_{n,i} (\boldsymbol{u}_i - \boldsymbol{u}_n) + 2\rho_n (\boldsymbol{u} \times \boldsymbol{\Omega}_E)$$

where

ρ_n	=mass density	$\boldsymbol{g}$	=Earth's gravitational acceleration
$\boldsymbol{u}_n$	=wind velocity	$\boldsymbol{\Omega}_E$	=Earth's angular velocity
p_n	=pressure	$\nu^*_{n,i}$	=momentum transfer collision frequency
η	=coefficient of viscosity	$\boldsymbol{u}_i$	=ion drift velocity

The subscripts n and i refer to the neutral and ionized components of the upper atmosphere, respectively. Evidently, this equation describes an equilibrium situation in which the inertial forces due to temporal and spatial changes in the wind velocity field are just balanced by the pressure gradient force, the viscosity force, the gravity force, the frictional force, and the Coriolis force. Besides the wind velocity $\boldsymbol{u}_n$, which is of primary interest here, this equation contains a number of variables which must either be eliminated using, for example, the continuity equation and the ideal gas law, or which must be specified beforehand. For example, the pressure gradient force may

be fixed using an empirical model of the disturbed temperature or by solving self-consistently the energy balance equation. Likewise, the ion drift velocity $\boldsymbol{u}_i$ and the momentum transfer collision frequency $\nu^*_{n,i}$ may be specified with the help of an ionospheric model. Given all this input information, the equation of motion may be solved using standard numerical methods. Simulations of this kind do indeed confirm that a sudden energy addition at high latitudes will launch a pulse-like superposition of atmospheric gravity waves which is associated with meridional winds of moderate magnitude and which moves with high velocity towards lower latitudes (e.g. Testud et al., 1975 [67]; Richmond and Matsushita, 1975 [59]; Chang and St. Maurice, 1991 [11]).

In a similar way, the ionospheric response to such thermospheric perturbations may be determined by solving the continuity equation for a charged particle species s

$$\frac{\partial n_s}{\partial t} = q_s - l_s - \boldsymbol{\nabla}(n_s \boldsymbol{u}_s)$$

As indicated by this equation, changes in the ionization density n_s may be caused by the production of ionization (q_s), by the loss of ionization (l_s), and by production or loss effects due to transport of ionization ($-\nabla(n_s \boldsymbol{u}_s)$). Again additional information is needed to solve this equation. For example, the solar EUV and X-ray radiation intensity and the associated absorption and ionization cross sections must be known. Also, numerous reaction rate constants are needed to properly describe the ion chemistry. In addition, the neutral gas densities and the neutral winds have to be known and are usually taken from empirical models of the upper atmosphere. Also, the temperatures of the different gas species, ionized or neutral, are needed. Finally, the ionization exchange with the plasmasphere has to be specified or calculated self-consistently. Given all this information, the continuity equation may be solved using standard numerical methods. This way various aspects of the ionospheric storm behaviour have been successfully simulated. Some of the more recent publications in this field include the studies by Bauske and Prölss (1998) [3], Mikhailov and Schlegel (1998) [41]), Richards and Wilkinson (1998) [57], Schlesier and Buonsanto (1999) [61], Mikhailov and Förster (1999) [40], Pavlov and Foster (2001) [45], Kilifarska and Ouzounov (2001) [34], Daniell and Strickland (2001) [15], Prölss and Werner (2002) [55], and references therein.

Another family of models exists which simulates both the neutral and ionized upper atmosphere self-consistently, sometimes even including the plasmasphere. In these models the non-linear balance equations of density, momentum, and energy are solved simultaneously. This way the densities of the neutral and ionized species, along with the neutral winds and the ionization drift, and the neutral, ion and electron temperatures are determined. Not surprisingly, these models are rather complex and sometimes difficult to validate. They also require large computing resources and have only limited spatial resolution. Finally, these models still depend on the proper specification of the energy input, which in the case of the solar wind energy source

is quite difficult. Nevertheless, these models have considerably matured over the last years and are now in a position to reproduce most of the salient features of upper atmospheric storms, even if only in a qualitative way (e.g. Burns et al., 1995 [10]; Fuller-Rowell et al., 1996 [25]; Fujiwara et al., 1996 [24]; Codrescu et al., 1997 [12]; Field et al., 1998 [21]; Förster et al., 1999 [22]; Emery et al., 1999 [18]; Namgaladze et al., 2000 [43]; Fuller-Rowell et al., 2000 [26]; Lu et al., 2001 [37]; and references therein). It is also clear that in the long run it is these models which will be used for space weather predictions, especially if data assimilation techniques are implemented.

Acknowledgements

I am very grateful to C. Berger who provided the CASTOR data used in this study. The DE data were kindly provided by the NASA National Space Science Data Center. I am grateful to all the experimenters who contributed to this data set, especially to G. Carignan, J. Craven and L. Frank. Thanks are also due to W. Keil who provided some information on the ASCA satellite. Last but not least, I am indebted to N. Ben Bekhti, K. Schrüfer and B. Winkel for their help in preparing this manuscript.

References

1. Anderson, C.N., *Proc. Inst. Radio Eng., 16*, 297-347, 1928
2. Anderson, C.N., *Proc. Inst. Radio Eng., 17*, 1528-1535, 1929
3. Bauske, R., and G.W. Prölss, *Adv. Space. Res., 22, No.1*, 117-121, 1998
4. Bauske, R., S.Noël, and G.W.Prölss, *Ann. Geophys., 15*, 300-305, 1997
5. Bencze, P., I. Almár, and E. Illés-Almár, *Adv. Space Res., 13, No.1*, 303-306, 1992
6. Berger, C., and F. Barlier, *J. Atmos. Terr. Phys., 43*, 121-133, 1981
7. Blumen, W., and R.G. Hendl, *J. Atmos. Sci., 26*, 210-217, 1969
8. Bremer, J., *Adv. Space Res., 22, No. 6*, 837-840, 1998
9. Burns, A.G., and T.L. Killeen, *Geophys. Res. Lett., 19*, 977-980, 1992
10. Burns, A.G., T.L. Killeen, W. Deng, G.R. Carignan, and R.G. Roble, *J. Geophys. Res., 100*, 14673-14691, 1995
11. Chang, C.A., and J.-P. St.-Maurice, *Can. J. Phys., 69*, 1007-1031, 1991
12. Codrescu, M.V., T.J. Fuller-Rowell, and I.S. Kutiev, *J. Geophys. Res., 102*, 14315-14329, 1997
13. Cole, K.D., *Nature, 194*, 75, 1962
14. Cole, K.D., and M.P. Hickey, *Adv. Space Res., 1, No. 12*, 65-75, 1981
15. Daniell Jr., R.E., and D.J. Strickland, *J. Geophys. Res., 106*, 30307-30313, 2001
16. Danilov, A.D., L.D. Morozova, Tc. Dachev, and I. Kutiev, *Adv. Space Res., 7, No.8*, 81-88, 1987

17. Dessler, A.J., *Nature, 184*, 261-262, 1959
18. Emery, B.A., C. Lathuillere, P.G. Richards, R.G. Roble, M.J. Buonsanto, D.J. Knipp, P. Wilkinson, D.P. Sipler, and R. Niciejewski, *J. Atmos. Solar-Terr. Phys., 61*, 329-350, 1999
19. Espenschiedt, L., C.N. Anderson, and A. Bailey, *Bell Syst. Techn. Journ., 4*, 459-507, 1925
20. Evans, J.V., *J. Geophys. Res., 75*, 4815-4823, 1970
21. Field, P.R., H. Rishbeth, R.J. Moffett, D.W. Idenden, T.J. Fuller-Rowell, G.H. Millward, and A.D. Aylward, *J. Atmos. Solar-Terr. Phys., 60*, 523-543, 1998
22. Förster, M., A.A. Namgaladze, and R.Y. Yurik, *Geophys. Res. Lett., 26*, 2625-2628, 1999
23. Foster, J.C., *J. Geophys. Res., 98*, 1675-1689, 1993
24. Fujiwara, H., S. Maeda, H. Fukunishi, T. J. Fuller-Rowell, and D. S. Evans, *J. Geophys. Res., 101*, 225-239, 1996
25. Fuller-Rowell, T.J., M.V. Codrescu, H. Rishbeth, R.J. Moffett, and S. Quegan, *J. Geophys. Res., 101*, 2343-2353, 1996
26. Fuller-Rowell, T.J., M.C. Codrescu, and P. Wilkinson, *Ann. Geophys., 18*, 766-781, 2000
27. Georges, T.M., *Ionospheric effects of atmospheric waves*, p. 3, *ESSA Techn. Rep. IER 57-ITSA 54*, Boulder/CO, 1967
28. Gold, T., Personal communication in Hines (1965) and Thome (1968), 1963
29. Hines, C.O., *J. Geophys. Res., 70*, 177-183, 1965
30. Humboldt, A.v., *Ann. Phys., 29*, 425-429, 1808
31. Humboldt, A.v., *Kosmos*, *Vol. 1*, p. 24, Cotta'scher Verlag, Stuttgart und Tübingen, 1845
32. Jakowski, N., A. Jungstand, K. Schlegel, H. Kohl, and K. Rinnert, *Can. J. Phys., 70*, 575-581, 1992
33. Johnson, F.S., *J. Geophys. Res., 65*, 2227-2232, 1960
34. Kilifarska, N.A., and D.P. Ouzounov, *J. Geophys. Res., 106*, 30415-30428, 2001
35. Klostermeyer, J., Thermospheric heating by atmospheric gravity waves, *J. Atmos. Terr. Phys., 35*, 2267-2275, 1973
36. Lauter, E.A., *Naturwissenschaften, 48*, 473-474, 1961
37. Lu, G., A.D. Richmond, R.G. Roble, and B.A. Emery, *J. Geophys. Res., 106*, 24493-24504, 2001
38. Maris, H.B., and E.O. Hulburt, *Proc. Inst. Radio Eng., 17*, 494-500, 1929
39. Mendillo, M., and J.A. Klobuchar, *An atlas of the mid-latitude F-region response to geomagnetic storms*, *Tech. Rept. 74-0065*, 267 pp., Air Force Cambridge Res. Lab., Cambridge, MA, 1974
40. Mikhailov, A.V., and M. Förster, *J. Atmos. Solar-Terr. Phys., 61*, 249-261, 1999
41. Mikhailov, A. V., and K. Schlegel, *Ann. Geophys., 16*, 602-608, 1998
42. Mikhailov, A.V., M.G. Skoblin, and M. Förster, *Ann. Geophys., 13*, 532-540, 1995

43. Namgaladze, A.A., M. Förster, and R.Y. Yurik, *Ann. Geophys., 18*, 461-477, 2000
44. Noël, S., and G.W. Prölss, *J. Geophys. Res., 98*, 17317-17325, 1993
45. Pavlov, A.V., and J.C. Foster, *J. Geophys. Res., 106*, 29051-29069, 2001
46. Pi, X., M. Mendillo, M.W. Fox, and D.N. Anderson, *J. Geophys. Res., 98*, 13 677-13 691, 1993
47. Pickard, G.W., *Proc. Inst. Radio Eng., 15*, 83-97, 1927a
48. Pickard, G.W., *Proc. Inst. Radio Eng., 15*, 749-766 1927b
49. Pickard, G.W., *Proc. Inst. Radio Eng., 15*, 1004-1012, 1927c
50. Prölss, G.W., *J. Geophys. Res., 87*, 5260-5266, 1982
51. Prölss, G.W., *J. Geophys. Res., 98*, 5981-5991, 1993
52. Prölss, G.W., *Ionospheric F-region storms*, in *Handbook of Atmospheric Electrodynamics, 2*, (H.Volland, ed.), 195-248, CRC Press / Boca Raton, 1995
53. Prölss, G.W., and J.D. Craven, *Adv. Space Res., 22, No. 1*, 129-134, 1998
54. Prölss, G.W., and M. Ocko, *Adv. Space Res., 26, No.1*, 131-135, 2000
55. Prölss, G.W., and S. Werner, *J. Geophys. Res., 107, No. A2*, 10.1029/2001 JA 900126, 2002
56. Prölss, G.W., K. Najita, and P.C. Yuen, *J. Atmos. Terr. Phys., 35*, 1889-1901, 1973
57. Richards, P.G., and P.J. Wilkinson, *J. Geophys. Res., 103*, 9373-9389, 1998
58. Richmond, A.D., *J. Geophys. Res., 84*, 5259-5266, 1979
59. Richmond, A.D., and S. Matsushita, *J. Geophys. Res., 80*, 2839-2850, 1975
60. Scherliess, L., and B.G. Fejer, *Geophys. Res. Lett., 25*, 1503-1506, 1998
61. Schlesier, A. C., and M. J. Buonsanto, *Geophys. Res. Lett., 26*, 2359-2362, 1999
62. Stephan, A.W., S. Chakrabarti, and D.M. Cotton, *Geophys. Res. Lett., 27*, 2865-2868, 2000
63. Strickland, D.J., R.E. Daniell, and J.D. Craven, *J. Geophys. Res., 106*, 21049-21062, 2001a
64. Strickland, D.J., J.D. Craven, and R.E. Daniell Jr., *J. Geophys. Res., 106*, 30291-30306, 2001b
65. Tanaka, T., *Geophys. Res. Lett., 13*, 1399-1402, 1986
66. Testud, J., *J. Atmos. Terr. Phys., 32*, 1793-1805, 1970
67. Testud, J., P. Amayenc, and M. Blanc, *J. Atmos. Terr. Phys., 37*, 989-1009, 1975
68. Thome, G., *J. Geophys. Res., 73*, 6319-6336, 1968
69. Tinsley, B.A., *J. Atmos. Terr. Phys., 43*, 617-632, 1981
70. Villain, J.P., *Ann. Geophys., 36*, 41, 1980
71. Volland, H., *J. Geophys. Res., 72*, 2831-2841, 1967
72. Volland, H., and H.G. Mayr, *J. Geophys. Res., 76*, 3764-3776, 1971
73. Wymore, I.J., *Proc. Inst. Radio Eng., 17*, 1206-1213, 1929

Space Weather Effects in the Upper Atmosphere: High Latitudes

Kristian Schlegel

Max Planck Institut für Sonnensystemforschung, 37191 Katlenburg-Lindau, Germany

Abstract. The most important space weather effects on the ionosphere and atmosphere, like the consequences of particle precipitation on conductivities and currents, are described. Following a description of related magnetic signatures on the ground and the relevant geomagnetic indices is a section on the Aurora as a prominent visible aspect of space weather. After a brief discussion of how space weather affects modern communication and navigation, the chapter concludes with an overview of solar flare and cosmic ray related effects.

1 Introduction

This chapter deals with the most important effects of space weather on the ionosphere and atmosphere. It is assumed that the reader is familiar with the basic physics and terminology of both "spheres". For reference in this text a figure is included showing the temperature, ion- and neutral density as a function of altitude (Fig. 1). Recent introductions into modern ionospheric physics are published by Kelley (1989) [9], Kohl et al. (1996) [11], Prölss (2001) [14], and Hagfors and Schlegel (2001) [7]. Many of the space weather effects can be summarized in a flow chart as displayed in Fig. 2. Processes in the magnetosphere, controlled by solar wind and described in previous chapters, cause two major phenomena in the upper atmosphere: particle precipitation and convection of the ionospheric plasma. Particle precipitation enhances the conductivity of the ionospheric plasma, while the plasma convection $\boldsymbol{V}$ in the presence of the Earth's magnetic field causes a system of electric fields due to

$$\boldsymbol{E} = -\boldsymbol{V} \times \boldsymbol{B} \tag{1}$$

Both, the enhanced conductivity σ and the electric field together cause large electric currents mainly in the auroral zone within the dynamo region (90–150 km altitude) according to Ohm's law

$$\boldsymbol{j} = \sigma \boldsymbol{E} \tag{2}$$

Consequences of these currents are Joule heating of the ionospheric plasma which is ultimately transferred to the neutral atmosphere, plasma instabilities, and observable changes of the geomagnetic field on the ground. Further

K. Schlegel, Space Weather Effects in the Upper Atmosphere: High Latitudes, Lect. Notes Phys. **656**, 215–238 (2005)
http://www.springerlink.com/

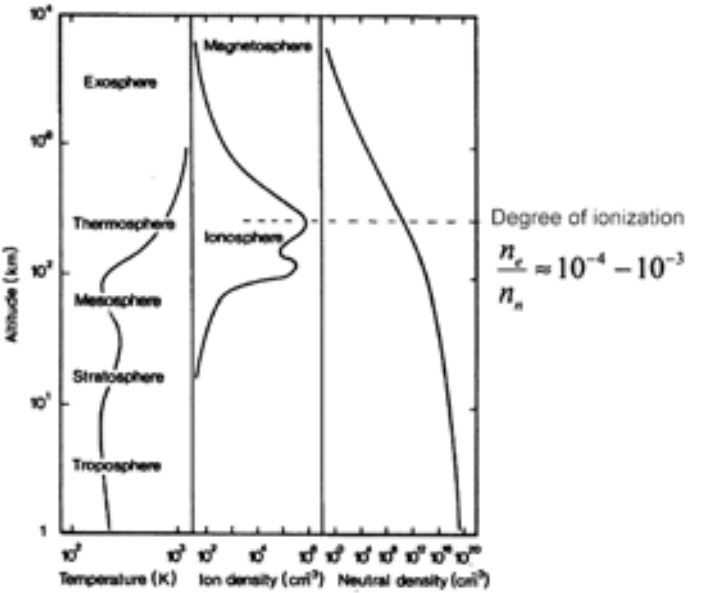


Fig. 1. Basic quantities of the atmosphere and ionosphere as a function of height.

consequences of particle precipitation are another type of heating of the ionospheric plasma and the aurora. The convection of the ionospheric plasma is also partly transferred to the neutral atmosphere by frictional processes and influences the global circulation of the neutral gas (see the chapter by G. Prölss, this volume).

In the following sections we will describe the processes sketched above in more detail.

2 Particle Precipitation

The particles precipitating from the magnetosphere into the ionosphere during space weather events are mainly electrons with energies of a few keV. They are spiralling around the geomagnetic field lines and interact with neutrals and ions by collisions. Neutrals are ionised by these collisions and the ion production as a function of particle energy and altitude can be described with the relation

$$\begin{aligned} q(E_p, z) &= \frac{FE_p}{E_{ion}} \Lambda(s/R) \frac{\rho(z)}{R(E_p)} \\ s &= \int_z^\infty \rho(h) dh \end{aligned} \tag{3}$$

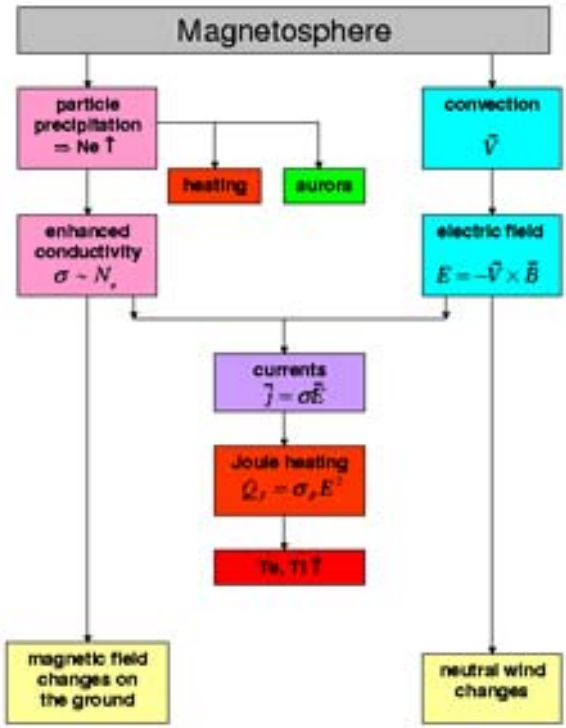


Fig. 2. Flow chart of space weather effects in the ionosphere and atmosphere.

F is the flux of the electrons, R their mean travel distance in air, E_p the peak of the electron energy spectrum, E_{ion} the mean ionisation energy of the neutral gas of 35 eV and $\rho(h)$ the density distribution of the neutral gas. The function Λ describes the energy dissipation of the electrons along their travel path s.

Typical ionisation rates as a function of altitude are plotted in Fig. 3. They are calculated under the assumption of mono-energetic electrons with the energy E_p. They are given for an electron flux of 10^8 particles/s/m^2/sterad and have therefore to be multiplied with the actual flux according to (3). They constitute a good approximation of the real case where the whole energy spectrum of the precipitating electrons has to be taken into account. Such calculations can only be performed in terms of computer simulations (e.g. Kirkwood and Osepian, 2001 [10]).

It is obvious from Fig. 3 that a peak electron energy of a few keV causes a maximum of ionisation in the lower E region. At high latitudes this ionisation can be more than an order of magnitude higher than the "usual" ionisation by solar EUV radiation. The resulting ion density N_i (= electron density N_e, because of charge neutrality) from the ionisation by particles Q_{Part} and solar

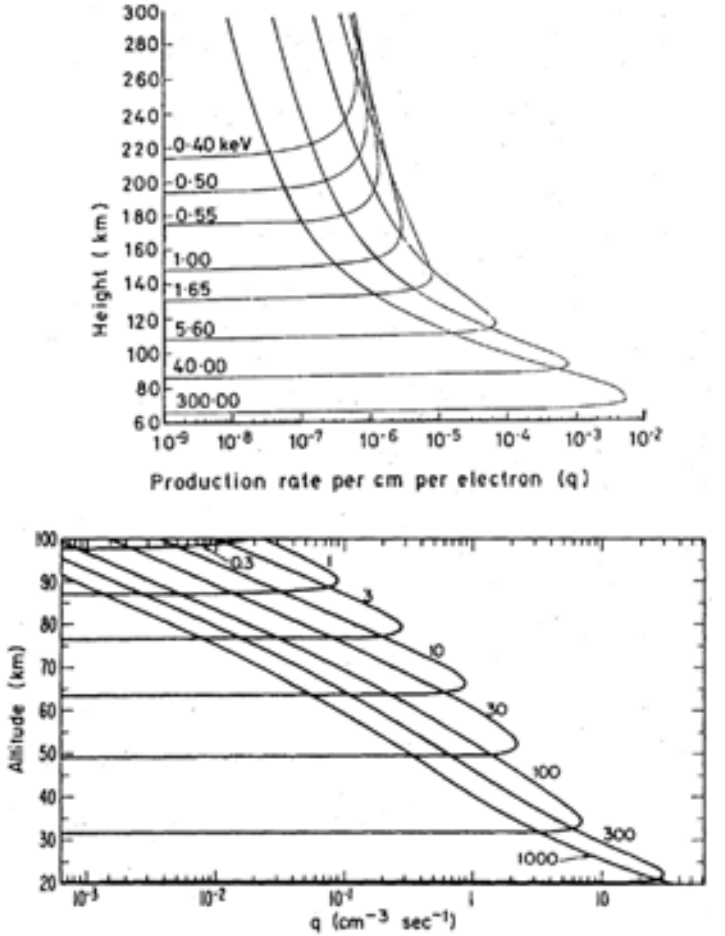


Fig. 3. Ionisation rates caused by precipitation particles (upper panel: electrons, lower panel: protons).

radiation Q_{EUV} can be calculated from the continuity equation

$$\frac{\partial N_i}{\partial t} = Q_{\mathrm{EUV}} + Q_{Part} - L - \mathrm{div}(N_i \boldsymbol{V}_i) \tag{4}$$

where L describes ionisation loss processes and the last term describes transport effects. Figure 4 shows an example of electron densities measured in the high latitude E region with the incoherent scatter technique (Alcaydé, 1997 [2]).

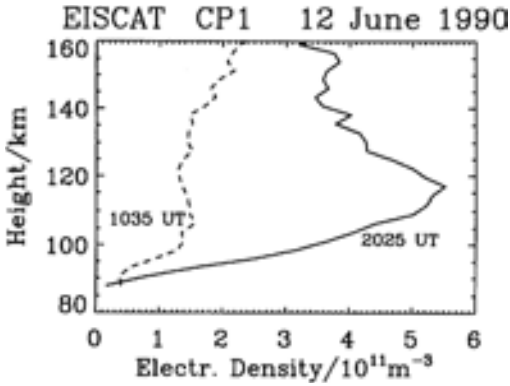


Fig. 4. Example of an electron density enhancement in the auroral E region during particle precipitation (at 20:25 UT) compared to quiet conditions (10:35 UT).

During solar proton events, a consequence of solar flares, high energy protons (up to more than 100 MeV) can precipitate into the ionosphere. Similar equations as (3) describe their ion production rate. Figure 3 also shows production rate profiles for mono-energetic protons. They are able to cause ionisation much deeper in the atmosphere than electrons (see Sect. 7).

3 Conductivities and Currents

Currents in the ionosphere are caused by moving charged particles, thus the current density is given by

$$\boldsymbol{j} = eN_e(\boldsymbol{V}_i - \boldsymbol{V}_e) \tag{5}$$

where e is the elementary charge, and V_i and V_e are ion and electron drifts, respectively. The latter can be calculated from the steady state momentum equations

$$\begin{aligned} &\text{ions}: && e(\boldsymbol{E} + \boldsymbol{V}_i \times \boldsymbol{B}) = m_i \nu_i n(\boldsymbol{V}_i - \boldsymbol{U}) \\ &\text{electrons}: && e(\boldsymbol{E} + \boldsymbol{V}_e \times \boldsymbol{B}) = m_e \nu_e n(\boldsymbol{V}_e - \boldsymbol{U}) \end{aligned} \tag{6}$$

where B is the geomagnetic field (in a coordinate system where $\boldsymbol{B} = (0, 0, B)$), m_i and m_e are ion and electron mass, ν_{in} and ν_{en} the ion-neutral and the electron-neutral collision frequency, respectively, and U the neutral wind.

Combining (5) and (6) yields for the current

$$\begin{aligned}\boldsymbol{j} &= eN_e(\boldsymbol{V}_i - \boldsymbol{V}_e)\\ &= \frac{eN_e}{B}\left\{\left(\frac{\omega_e\nu_{en}}{\omega_e^2+\nu_{en}^2}+\frac{\omega_i\nu_{in}}{\omega_i^2+\nu_{in}^2}\right)(\boldsymbol{E}+\boldsymbol{U}\times\boldsymbol{B})\right.\\ &\quad\left.+\left(\frac{\omega_e^2}{\omega_e^2+\nu_{en}^2}-\frac{\omega_i^2}{\omega_i^2+\nu_{in}^2}\right)(\boldsymbol{E}+\boldsymbol{U}\times\boldsymbol{B})\times\hat{b}\right\}\\ &\quad+e^2N_e\left(\frac{1}{m_e(\nu_{en}+\nu_{ei})}+\frac{1}{m_i\nu_i n}\right)E_\parallel \qquad (7)\\ \omega_{i,e} &= \frac{eB}{m_{i,e}}\end{aligned}$$

The three components of the current density vector are called Pedersen, Hall and parallel current. The Pedersen current flows perpendicular to B and parallel to E, the Hall current perpendicular to B and perpendicular to E, the latter is usually the strongest component. The conductivity can be expressed as a tensor $\tilde{\sigma}$, the current density thus becomes

$$\boldsymbol{j} = \tilde{\sigma}(\boldsymbol{E}+\boldsymbol{U}\times\boldsymbol{B}) \qquad (8)$$

with

$$\tilde{\sigma} = \begin{pmatrix}\sigma_P & \sigma_H & 0\\ -\sigma_H & \sigma_P & 0\\ 0 & 0 & \sigma_\parallel\end{pmatrix} \qquad (9)$$

The three components of this tensor are consequently

$$\begin{aligned}\text{Pedersen conductivity}: \quad \sigma_P &= \frac{eN_e}{B}\left(\frac{\omega_e\nu_{en}}{\omega_e^2+\nu_{en}^2}+\frac{\omega_i\nu_{in}}{\omega_i^2+\nu_{in}^2}\right)\\ \text{Hall conductivity}: \quad \sigma_H &= \frac{eN_e}{B}\left(\frac{\omega_e^2}{\omega_e^2+\nu_{en}^2}-\frac{\omega_i^2}{\omega_i^2+\nu_{in}^2}\right) \qquad (10)\\ \text{parallel conductivity}: \quad \sigma_\parallel &= e^2N_e\left(\frac{1}{m_e\left(\nu_{en}+\nu_{ei}\right)}+\frac{1}{m_i\nu_{in}}\right)\end{aligned}$$

Two important facts can be derived from these equations: (i) all conductivities are proportional to electron density which means that enhanced electron densities as caused by particle precipitation, in turn lead to high conductivities and currents, (ii) all conductivities are strongly height-dependent, because the collision frequencies ν_{in} and ν_{en} are proportional to the neutral density which decreases with altitude (Fig. 1). The latter is usually taken from a neutral atmospheric model (*http://nssdc.gsfc.nasa.gov/space/model/models/msis_n.html*).

Typical quiet-time conductivities are displayed in Fig. 5. Note that σ_H has its peak around 100 km altitude and decreases rapidly with height, whereas

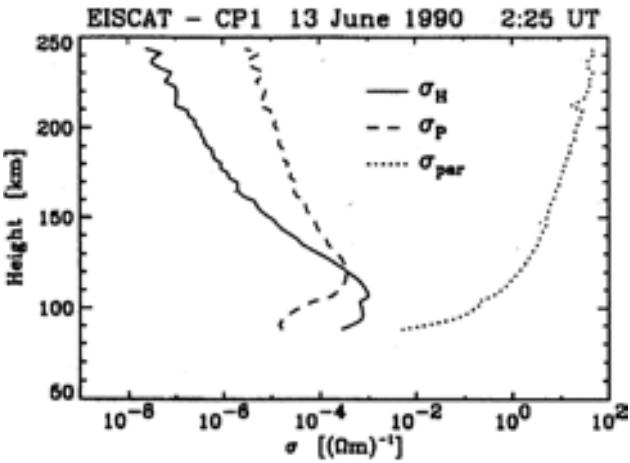


Fig. 5. Typical ionospheric conductivities as derived from EISCAT measurements (Kelley, 1989 [9]).

σ_P has its peak at about 120 km altitude and decreases slowly with height. Although the parallel conductivity is the highest of all three, since the charged particles can move freely along the magnetic field lines, it does not play any role because the parallel electric field is negligibly small. Thus the third term of the current density (7) is not important in the ionosphere.

For practical purposes often the so-called conductance is used. It is the conductivity integrated over the height range where it is most important, i.e.

$$\Sigma_{P,H} = \int_{90\,\mathrm{km}}^{250\,\mathrm{km}} \sigma_{P,H}\,(z)\,dz\ . \tag{11}$$

It has the advantage that it can be conveniently plotted versus time, thus showing the ionospheric variability during space weather events. Figure 6 gives an example which has been computed from measured electron densities (incoherent scatter technique) together with model values of collision frequencies (e.g. Schlegel, 1988 [20]). During the morning and afternoon of that day both, σ_H and σ_P are low, corresponding to quiet conditions where the E region electron density is mainly caused by solar EUV radiation. In the evening and persisting into the next day burst-like conductance enhancements were observed during a geomagnetic storm.

Electric fields in the ionosphere are low during quiet conditions, typically a few mV/m. Generally this field can be regarded as independent of height within the range of interest (90 to several 100 km). During strong geomagnetic storms the field can well exceed 100 mV/m. Figure 7 shows as example the N–S and the E–W components of the electric field during the same time interval as in Fig. 6. Large northward electric fields are observed in the so-

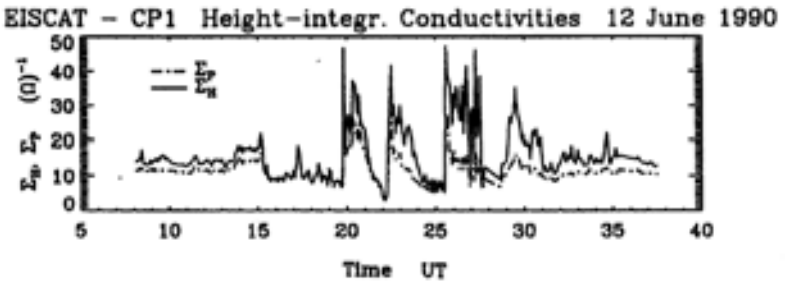


Fig. 6. Hall and Pedersen conductances during a magnetic storm.

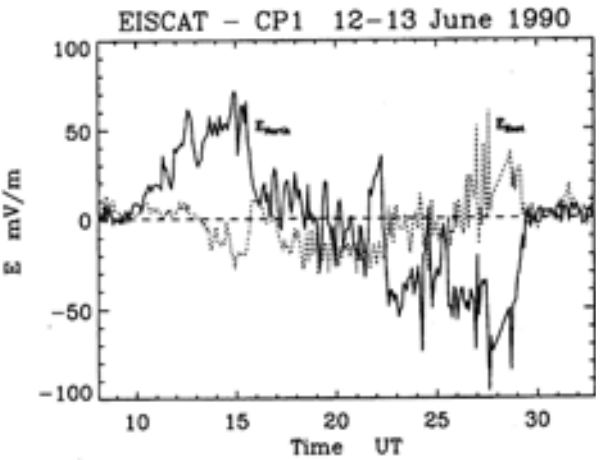


Fig. 7. Typical electric field variations during a magnetic storm.

called afternoon sector, the time between about 11:00 and 17:00 UT. The field turns southward in the subsequent morning sector after about 19:00 UT and remains in this direction until early next morning. This is a very typical electric field behaviour during a magnetic storm. The bursts-like enhancements in the morning sector correspond to substorms.

A typical current density profile during disturbed conditions in the morning sector is plotted in Fig. 8. The peak current flows in a relatively narrow height range between about 90 and 130 km altitude. This current is called the auroral electrojet. It is eastwards directed in the afternoon sector and westwards in the morning sector. The north-south extend of the electrojet is typically about 100 km. Therefore the total current of the electrojet at 1:40 UT on 13 June 1990 was of the order of

$$J = 75\,\mu\mathrm{A/m}^2 \cdot 30000\,\mathrm{m} \cdot 100000\,\mathrm{m} = 0.23 \cdot 10^6\,\mathrm{A}\ . \qquad (12)$$

During very strong magnetic storms it can easily exceed 1 Million A.

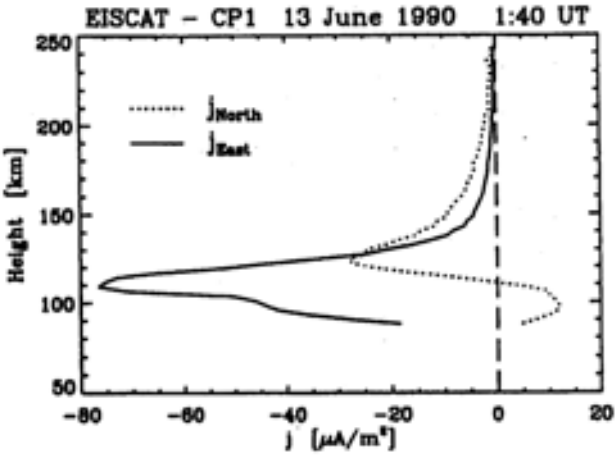


Fig. 8. Current densities as a function of height as derived from EISCAT data during particle precipitation and high electric fields.

The power (per unit volume) dissipated in the E region (Joule heating) by the electrojet is given by

$$P_j = \boldsymbol{j} \cdot \boldsymbol{E} = \sigma_P E^2 \ . \tag{13}$$

The product of the expression for the current (7) and the electric field vector reveals that the Hall current does not contribute to the power, it is a blind current. Microscopically the Joule heating arises from the friction between the charged particles and the neutrals (ν_{in} and ν_{en}). The height-integrated Joule heating

$$Q_J = \Sigma_P \, E^2 \tag{14}$$

can again conveniently be plotted versus time and gives thus an overview of the dissipated energy during a magnetic storm. Figure 9 shows an example which additionally contains the energy input from precipitating particles which can be estimated with the relation

$$Q_P = 1/2 \, \eta \alpha_{eff} N^2 \tag{15}$$

with $\eta = 35\,\mathrm{eV}$ and α_{eff} the effective electron-ion recombination rate. This quantity is generally smaller than the Joule heating, except during brief bursts of precipitation.

Q_J and Q_P constitute the main energy sinks during a magnetic storm, i.e. the energy transferred to the terrestrial atmosphere in a space weather event. It is therefore interesting to compare this energy with the total energy transferred to the magnetosphere by the solar wind. During the geomagnetic storm of 10 January 1997 the energy dissipated by Joule heating over the

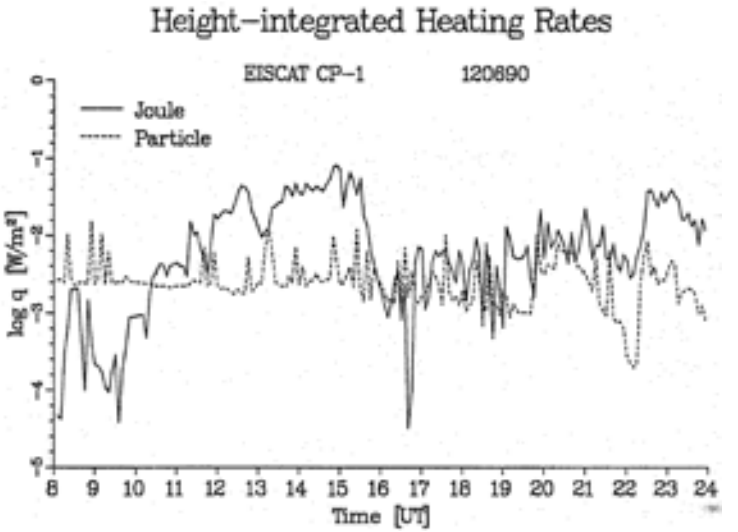


Fig. 9. Joule and particle heating in the auroral ionosphere during a magnetic storm.

whole auroral zone was estimated to 13000 TJ which is about 40% of the average solar wind energy input into the magnetosphere (Schlegel and Collis, 1999 [21])).

The energy transferred to the atmosphere causes first of all the ion and electron gas to be heated, but ultimately this energy is passed to the neutral gas. It causes a considerable expansion of the auroral atmosphere. This has important consequences for satellites orbiting the Earth at altitudes below about 500 km as already mentioned in the previous chapter by G. Prölss.

It should be noted in this context that a heating of the terrestrial atmosphere occurs regularly within the solar cycle, apart from magnetic storms. Whereas the visible and infrared part of the solar spectrum does only marginally change during the solar cycle, the EUV-flux in the wavelength range below 100 nm is increased by more than a factor of three during solar activity maximum years. This yields a higher energy input into the upper atmosphere, since this part of the solar radiation is mainly absorbed at altitudes above 100 km. Consequently not only the neutral gas density and temperature but also the ionisation is increased, as demonstrated in Fig. 10. Apart from the consequences for satellite trajectories this also leads to important differences in short wave radio propagation during the solar cycle, as known for more that 70 years. Even radio amateurs enjoy the greater distances to be covered during solar maximum years.

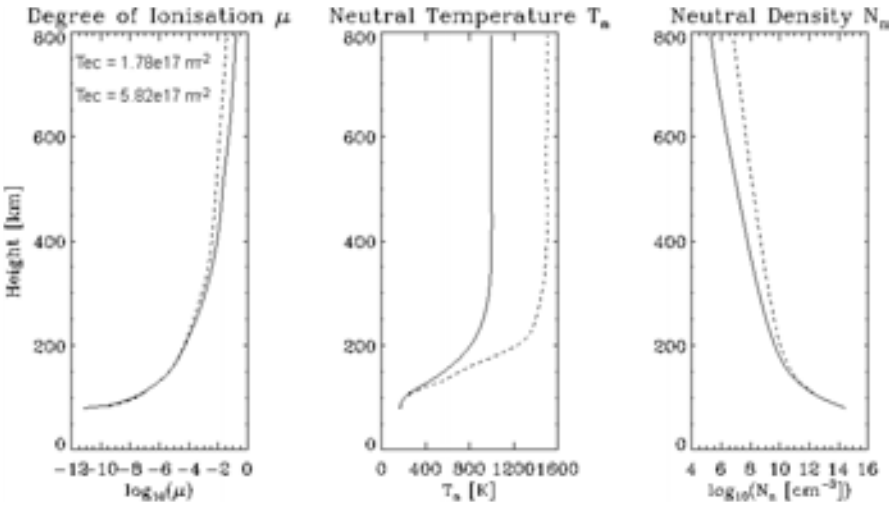


Fig. 10. Difference between high and low solar activity in atmospheric ionisation, temperature, and density.

4 Magnetic Signatures on the Ground and Geomagnetic Indices

The auroral electrojet causes distinct perturbations of the geomagnetic field which can be monitored with magnetometers on the ground. Although the Hall current is a blind current, it usually causes the strongest variations $\Delta \boldsymbol{B}$. According to the "right-hand-rule" the magnetic perturbation appears mainly in the N–component of $\Delta \boldsymbol{B}$ in the evening sector and in the S–component in the morning sector. With the help of N–S aligned magnetometer chains the location and extend of the electrojet can be well established, Fig. 11 shows an example.

Equally important are ground-based magnetometers for the derivation of geomagnetic indices which are widely used to characterize space weather events in a quantitative manner. Since the pioneering work of the German geophysicist Julius Bartels (1899–1964) geomagnetic storms are characterised by the index Kp (Chapman and Bartels, 1962 [4]). Bartels who introduced this index in 1949 derived it from the largest variation of the horizontal magnetic field component during a 3–hour interval from a single magnetometer station, using a quasi-logarithmic scale. This so-called K index was then averaged over 13 globally distributed stations, applying special weighting functions, in order to obtain the Kp index where p stands for "planetary". Kp runs from 0 (very quiet) to 9 (very disturbed) and is further subdivided using the subscripts $-, 0, +$ (e.g. $1-, 1_0, 1+, 2-, ...$), this yields 28 steps in total. Bartels also developed a convenient representation of Kp in terms of musical notes (Fig. 12).

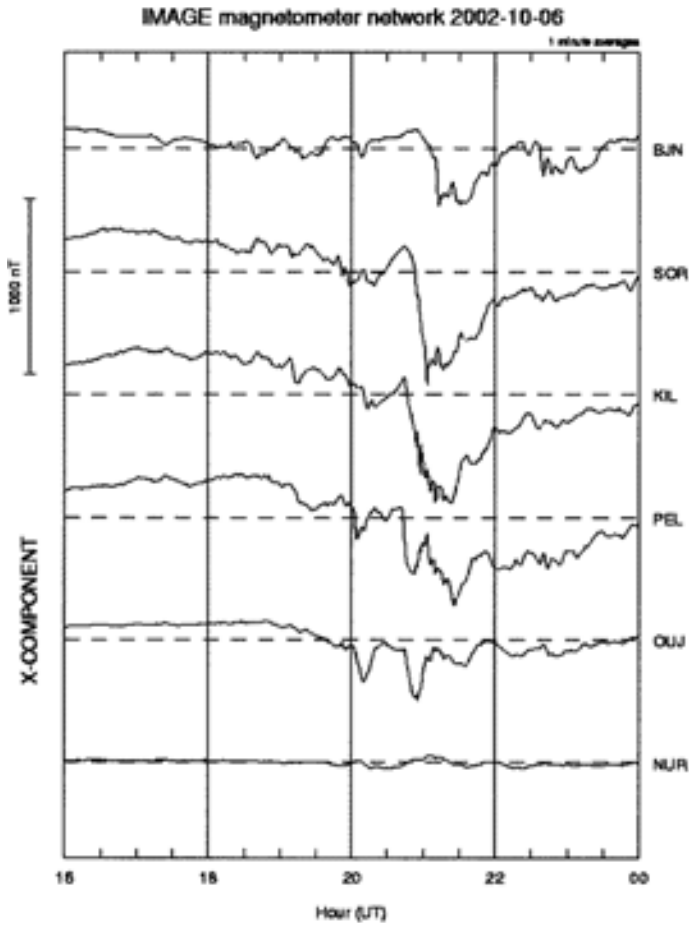


Fig. 11. Record of the x-component of the geomagnetic field from a part of the IMAGE magnetometer network. The instruments are apprximately meridionally aligned from southern Finland (NUR) to Bear Island (BJN).

As already mentioned, Kp is expressed in a logarithmic scale and consequently not very well suited for averaging. Bartels therefore introduced the linear equivalent ap where $\mathrm{Kp} = 9_0$ corresponds to $\mathrm{ap} = 400\,\mathrm{nT}$. The Ap index is a mean over eight 3–hour intervals of ap, i.e. over a full day. It consequently characterises not only the strength but also the duration of the strongest phase of a storm.

Bartels was able to derive both indices back to 1932, for earlier years not enough magnetometer stations were available. For many years the University of Göttingen issued the Kp and Ap indices, but since beginning of 1997 this task has been taken over by the Adolf-Schmidt Observatorium für Ge-

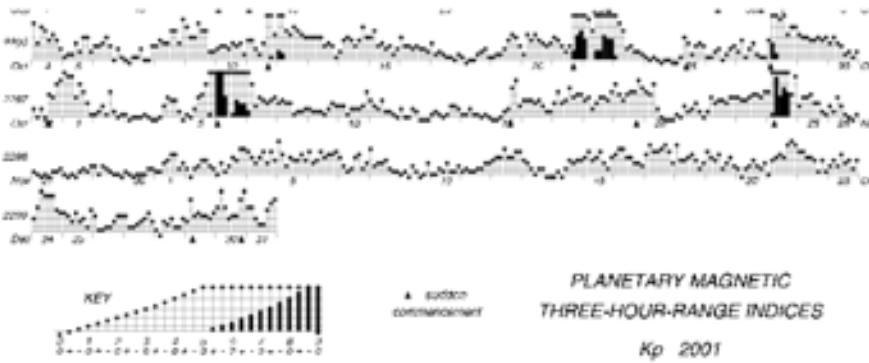


Fig. 12. Bartels' Kp-notation as musical notes. Every line represents 28 days, the average solar rotation. Recurrency trends can therefore be easily detected.

omagnetismus in Niemegk/Germany (*http://www.gfz-potsdam.de/pb2/pb23/GeoMag/niemegk/obs_eng.html*).

In order to characterize geomagnetic storms before 1932 a different index, the so-called AA index was developed. It is similar as Ap but is derived from the magnetograms of only two stations, one on the northern (England) and one on the southern hemisphere (Australia). Since both stations have recorded the geomagnetic field since 1868, it was possible to derive aa (3–h interval) and AA (full day) back to this year. Finnish scientist have recently pushed the AA-records even further back (Nevanlinna and Kataja, 1993 [13]).

The so far mentioned indices characterise geomagnetic variations particularly at high and midlatitudes which are mainly related with the auroral electrojet. The magnetic variations due to the ring current (see Chapter by G. Prölss) are described by the Dst index. It is derived since 1957 from the horizontal magnetic field component measured at 4 stations near the equator. The magnetic field of the ring current is directed opposite to the main geomagnetic field, consequently strong disturbances are characterised by large negative Dst excursions.

Finally, magnetic disturbances at very high latitudes are characterised by the AE index which is derived from magnetic records of 12 stations at auroral latitudes. Details of the derivation of all indices can be found in Mayaud (1080) [12], their values are accessible through the internet (*http://www.cetp.ipsl.fr/~isgi/homepag1.htm, http://spidr.ngdc.noaa.gov/spidr/html*). A map with the stations for the various indices is given in Fig. 13. The convenience of geomagnetic indices is demonstrated with Table 1, listing the 10 strongest storms of the past century.

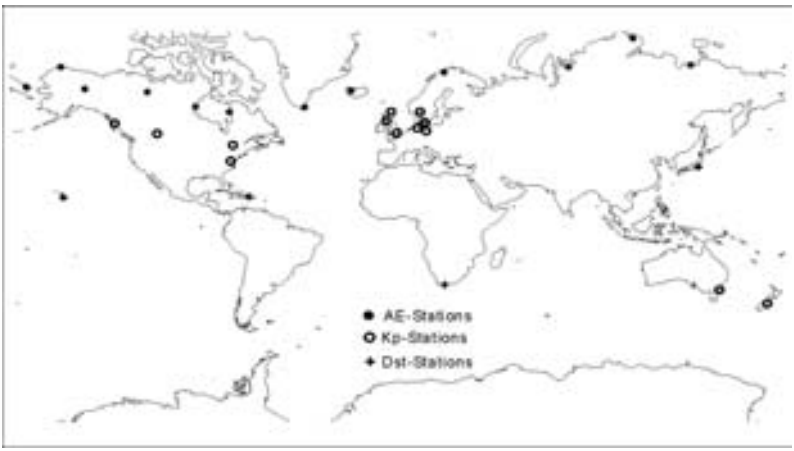


Fig. 13. Map of the location of magnetometer stations from which the various magnetic indices are derived.

Table 1. The 10 strongest geomagnetic storms in the past century in descending order. The second column gives AA*= Σaa of the most strongest 24–h interval (AA, without the star corresponds to one day.) The third column gives the maximal Kp (since 1932) and the fourth the minimal Dst (since1957) in the corresponding interval. The last column shows the location of auroral observations most near to the equator during the storms. Si refers to a list of auroral observations compiled by Silverman ranging from 686 BC to 1951 AD, Sch to W. Schröder (private communication), A to other sources.

Date	AA* [nT]	max. Kp	min. Dst [nT]	Auroral Observation nearest to the Equator (geogr. Latitude)
1989 13./14. March	441	9_0	–589	A: Florida Keys, ($\Phi \approx 24^\circ$ N)
1941 18./19. Sept.	429	9–	–	Si: Florida, ($\Phi \approx 29^\circ$ N)
1940 24./25. March	377	9_0	–	Si: Korfu, ($\Phi = 39^\circ$ N)
1960 12./13. Nov.	372	9_0	–339	A: Atlantic, ($\Phi = 28^\circ$ N)
1959 15./16. July	357	9_0	–429	Sch: $\Phi \approx 48^\circ$ N
1921 14./15. May	356	–	–	Si: Samoa, ($\Phi = 14^\circ$ S)
1909 25./26. Sept.	333	–	–	Si: Mallorca, ($\Phi = 39^\circ$ N)
1946 28./29. March	329	9_0	–	Si: Queensland, ($\Phi \approx 27^\circ$ S)
1928 7./8. July	325	–	–	Si: Atlantic, ($\Phi = 24^\circ$ N)
1903 31.10./1.11.	324	–	–	Si: Bamberg, ($\Phi = 50^\circ$ N)

5 Aurora

Aurora is the only visible and pleasant aspect of space weather. They are caused by the aforementioned keV-particles precipitating from the magnetosphere into the upper atmosphere. At altitudes between about 500 and 90 km these particles interact with atmospheric constituents, mainly N_2, O_2 and O. These constituents are excited and subsequently radiate the excitation energy

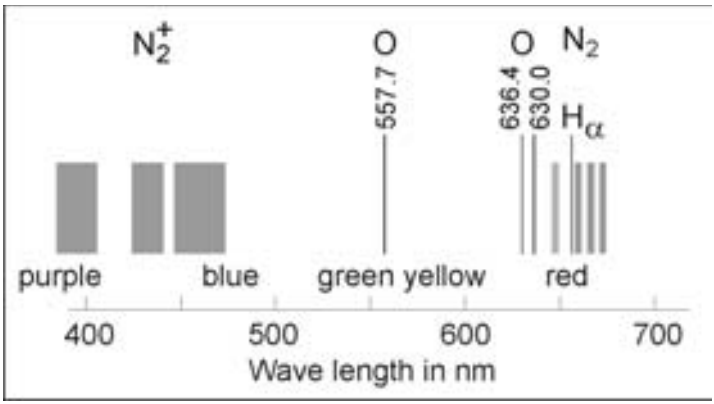


Fig. 14. Simplified spectrum of auroral emissions in the visible range.

over a broad spectrum (infrared, visible, ultraviolet). It should be noted that only a small part of the emissions are caused by direct collisional excitation through the precipitating particles or their secondaries, the major part is released in chemical reactions which are in turn induced or affected by these particles. Figure 14 shows a simplified spectrum of auroral emissions in the visible range.

The predominant green colour of aurora is caused by the following reactions

$$\mathrm{N}^+ + \mathrm{OH}^2 \quad \rightarrow \quad \mathrm{NOH}^+ + \mathrm{OH}\ \left(^1\mathrm{S}\right),$$

the excited oxygen atom then transits in a lower excitation state by emitting a photon

$$\mathrm{O}\ \left(^1\mathrm{S}\right) \quad \rightarrow \quad \mathrm{O}\ \left(^1\mathrm{D}\right) + \mathrm{Photon}\,(555.7\,\mathrm{nm}).$$

Red light is emitted when this metastable state goes to the ground state

$$\mathrm{O}\ \left(^1\mathrm{D}\right) \quad \rightarrow \quad \mathrm{O}\ \left(^3\mathrm{P}\right) + \mathrm{Photon}\,(630\,\mathrm{nm}).$$

The O $\left(^1\mathrm{D}\right)$-state can also directly be excited by dissociative recombination:

$$\mathrm{O}_2^+ + e \quad \rightarrow \quad \mathrm{O} + \mathrm{O}\ \left(^1\mathrm{D}\right)$$

Nitrogen molecules can be ionised and excited by primary electrons:

$$\mathrm{N}_2 + \mathrm{e}_{\mathrm{primary}} \quad \rightarrow \quad \mathrm{N}_2^+\left(\mathrm{B}^2\Sigma_u^+\right) + \mathrm{e}_{\mathrm{primary}} + \mathrm{e}_{\mathrm{sekundary}}$$

This excited state of the ionised nitrogen molecule is a vibrational state. In the transition to other vibration states a whole band of colours in the blue-violet range is emitted. Emissions of the neutral nitrogen molecule fall within the red and ultraviolet bands (mode details of emissions, see Rees, 1998 [16]).

The green and the red line (the latter is actual a doublet) are so-called forbidden lines. The corresponding excitation states have a relatively long

Table 2.

IBC	Intensity of the 557.7 nm line (kR = kilo Rayleigh)	Comparable brightness
I	1 kR	as the milky way
II	10 kR	moonlight on thin cirrus
III	100 kR	moonlight on cumulus
IV	1000 kR	full moon

life time of 1 s (green) and 110 s (red). Under normal pressure at ground level these excited states would be immediately quenched by collisions with other atmospheric constituents. Only at altitudes above 100 km and pressures below 0.1 Pa, the mean time between two collisions is longer than the excitation life time and the de-excitation by emission becomes possible. The association of these lines to atomic oxygen was therefore a longstanding problem to spectroscopists and was finally solved not before 1932.

The brightness of aurora is characterized by the "international brightness coefficient" (IBC) according to four classes (1 Rayleigh = 106 photons/cm^2/s/sterad) as listed in Table 2.

The special topology of the geomagnetic field lines extending into the magnetospheric tail cause the aurora to be confined mainly to a ring around the magnetic poles, the so-called auroral oval (Fig. 15). Within this ring which is located at about 70° geomagnetic latitude and has a typical width of several 100 km during not too disturbed conditions, aurora occurs most frequently. During very strong space weather events the auroral oval expands towards the equator and can easily reach mid latitudes. Due to the smaller dip angle of the field lines the auroral particles experience a longer travel time through the atmosphere and therefore aurora appears mainly at altitudes above 200 km as a red glow. These red colours were associated with blood by our ancestors, and therefore aurora was regarded as a bad omen for war and diseases (Schlegel, 2001 [18]). The forms of aurora depend on the topology of the currents flowing from the magnetotail into the polar regions. In principle two basic manifestations exist: diffuse aurora (unstructured, extended) and discrete aurora (arcs, veils, bands, localised). Some common forms are sketched in Fig. 16. The diffuse aurora is caused by particles in the 100 eV range which are scattered into the loss cone, and appear mainly at altitudes above 150 km. The energy of particles causing the discrete aurora on the other hand, is of the order of several keV as already mentioned, considerably lower than the mean energy of the particles in the plasma sheet of the tail, their origin. The particles have therefore to be accelerated along the field lines. The nature of this acceleration is still under debate, several different mechanisms are discussed (e.g. Schlegel, 1991 [19]).

Auroral particles causing the so far mentioned aurora are electrons. The "proton aurora" is much more rare and is caused by energetic protons which

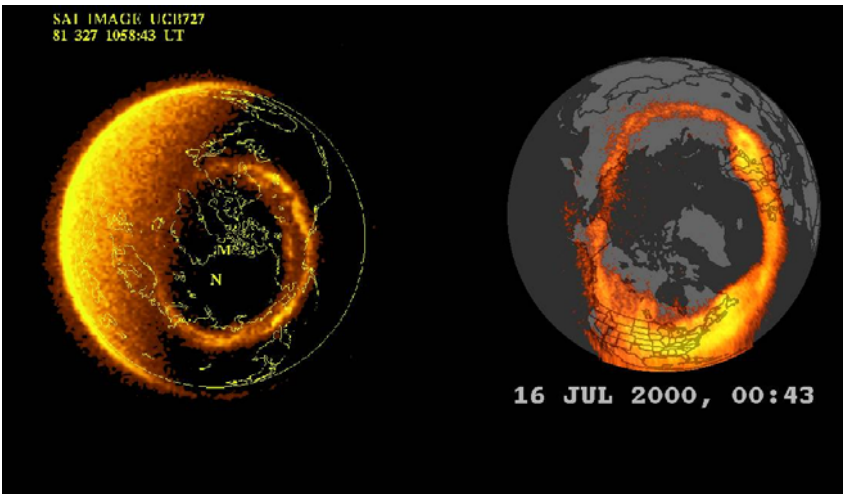


Fig. 15. Auroral oval during quiet (left) and disturbed (right) conditions.

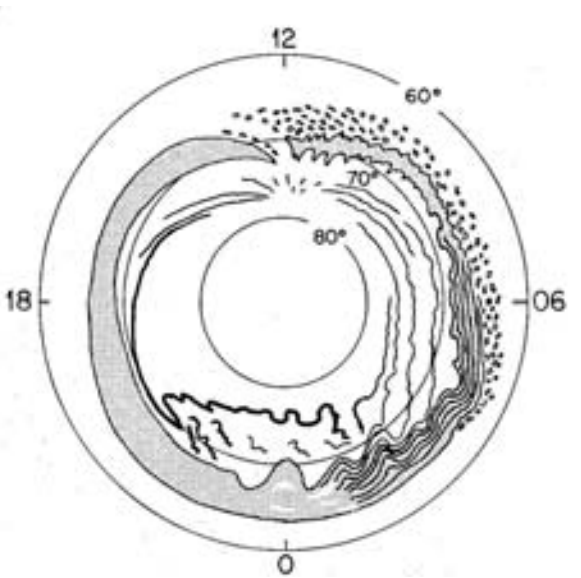


Fig. 16. Schematic representation of the main auroral forms as a function of local time for latitudes $> 60°$. The shaded area characterises diffuse aurora, the thick line a quiet arc which transforms into folded bands after about 21:00 LT. In the morning hours patchy aurora can often be found at the southern rim of the auroral oval. The short thin lines around local noon at about 75° latitude are daylight aurora (Akasofu, 1970 [1]).

Fig. 17. Four auroral displays with different colours and forms. The lower right one was photographed at mid-latitudes (near Düsseldorf).

are decelerated in the atmosphere by collisions and finally transformed to excited neutral hydrogen by charge transfer:

$$M \ + \ \mathrm{H} \quad \rightarrow \quad M^{+} \ + \ \mathrm{H}^{*},$$

where M is any neutral constituent. The excited hydrogen atoms emit L_α (121.57 nm, UV) or H_α (656.3 nm, red). The latter cannot be distinguished by eye from the red oxygen light (Fig. 14). Proton aurora is generally diffuse and often associated with PCA events (Sect. 7).

There are plenty of internet pages with splendid auroral photos, e.g. *http://www.meteoros.de, http://www.exploratorium.edu/learning_studio/auroras/, http://sgo.fi/Pictures/, http://www.pi.physics.uiowa.edu/vis/*. A brief collection is printed in Fig. 17.

6 Consequences of Electron Density Enhancements and Fluctuations

The enhancement of electron density by precipitating particles as described in Sect. 2 has important consequences on communication and navigation.

Although the importance of HF communication which is most strongly affected, has decreased in recent years, it still plays a role in many countries. It is therefore necessary to forecast possible changes of HF propagation during space weather events.

The propagation of electromagnetic waves in the ionosphere is described by the magneto-ionic theory (Rawer, 1993 [15]). One important equation is the index of refraction of the waves which in its simplest form reads

$$n^2 = 1 - \frac{\omega_P^2}{\omega^2} \tag{17}$$

$$\text{with the plasma frequency } \omega_P = \sqrt{\frac{N_e e^2}{\epsilon_0 m_e}} \ .$$

It is obvious from this equation that the propagation of a wave with frequency ω depends strongly on the plasma frequency and thus on the electron density. Waves used for communication under quiet conditions may not reach their destination (for instance, when n becomes imaginary) under conditions with enhanced electron density.

Eq uation (16) indicates a strong decrease of ionospheric propagation effects for large frequencies, for $\omega \gg \omega_P$, the refractive index approaches unity which means propagation in vacuum space. But even for GHz radio waves the propagation effects are not negligible in certain cases, for instance in GPS navigation.

A very important ionospheric quantity in this context is the total electron content

$$TEC = \int_P N_e \ ds \ , \tag{18}$$

where the integral is taken over the signal path from the ground station to the satellite. Typical values are of the order of 50–150 TECU (TEC-units, 10^{16} electrons/cm^2). Due to enhanced electron densities during particle precipitation changes of the order of several 10 TECU can easily occur. Figure 18 shows an example. In case of GPS measured distances the TEC change translates into errors of

$$D(\text{mm}) = 2.16\, TEC \text{ (in TECU)} \tag{19}$$

It should be noted that electron density enhancements may not only occur at high latitudes, but can also be convected towards lower latitudes as so called patches.

The strong currents in the auroral E-region cause plasma instabilities which lead to a structuring of the normally uniform plasma. The formed plasma irregularities have a broad range of scale lengths λ_{irr}, from kilometres

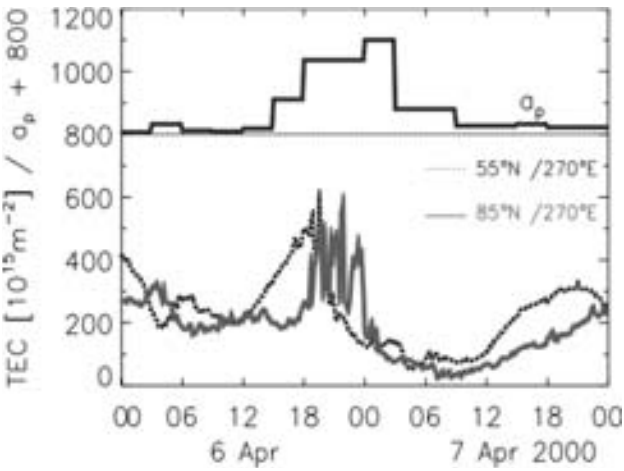


Fig. 18. Variation of TEC during a magnetic storm, the upper panel shows ap (Jakowski et al., 2002 [6]).

to tenth of meters and can therefore cause constructive interference with radio waves. This can lead either to strong backscatter or to forward scatter of radio waves in a wide frequency band whenever

$$\lambda_{radio\ wave} = 2\lambda_{irr} \tag{20}$$

A corresponding effect which is often observed during space weather events is the overrange of VHf signals, e.g. that taxi drivers in Hamburg can listen to their colleagues in Helsinki over their usual communication channels. Radio amateurs too use this "auroral scatter" as they termed it, for long range communication.

Even satellite signals in the GHz range are affected in such cases. This "radio scintillation" causes amplitude and phase fluctuations of satellite signals and thereby disturbs the communication and also degrades the accuracy of GPS measurements (Basu and Groves, 2001 [3]).

7 Solar Flare and Cosmic Ray Related Effects

As mentioned in previous chapters during and after a solar flare the flux of high energy protons as well as of X-rays is enhanced at the Earth by several orders of magnitude. During the very strong flare on 18 August 1979 for instance, the X-ray flux in the wavelength range 0.029–0.048 nm increased by a factor of 2000, and that of the 0.05–0.8 nm by a factor of 280. Solar X-rays play in general an important role in the ionisation of the ionospheric D-region. An enhancement of their flux can therefore considerably increase the electron density in the height range 80–100 km (Collis and Rietveld, 1990

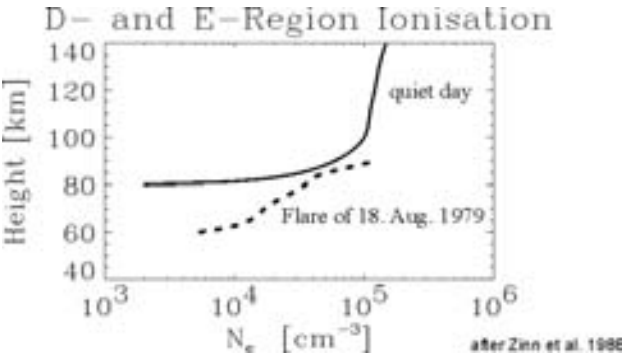


Fig. 19. Electron density during quiet conditions and during the solar flare of 18 Aug. 1979.

[5]). A similar ionisation increase cause the high energy protons which can penetrate well down into the stratosphere (see Fig. 3 above). Figure 19 shows an example of the *D*-region electron density increase during the above mentioned flare. Whereas the X-rays reach the Earth only about 8 min after the flare onset, the energetic protons need a travel time of the order of one hour. The X-ray flux increase is peak-like with a duration of only about 10 min, whereas the enhanced proton flux usually pertains for several days. Thus the large electron densities in the mesosphere and stratosphere maintain for a similar time. High electron densities together with the high electron-neutral collision frequencies at *D*-region heights cause a strong damping of electromagnetic waves according to magneto-ionic theory (Rawer, 1993 [15]). Thus short (MHz) and medium (kHz) wave communication is strongly affected in such cases.

In the vicinity of the Earth the energetic protons gyrate around the geomagnetic field lines according to Störmer's theory (e.g. Walt, 1994 [22]). An important quantity for their propagation is the "magnetic rigidity"

$$R = \frac{pc}{Ze} \ , \tag{21}$$

where p is the particle momentum, c the velocity of light and Z their charge number (this formula also applies to particle with $Z > 1$, e.g. alpha particles). All particles with the same rigidity have the same orbit parameters. It can be shown that all particles with a critical rigidity

$$R_c = 14.9 \cos^4 \lambda_c \tag{22}$$

reach geomagnetic latitudes $\lambda \geq \lambda_c$, or differently expressed, all particles with $R \geq R_c$ can reach the geomagnetic latitude λ_c. This is explained in Fig. 20: protons with energies $E_p < 100$ MeV will penetrate the Earth's atmosphere only at high latitudes; the higher their energy, the more lower latitudes they can reach. Since the peak of solar protons is normally below 100 MeV, the

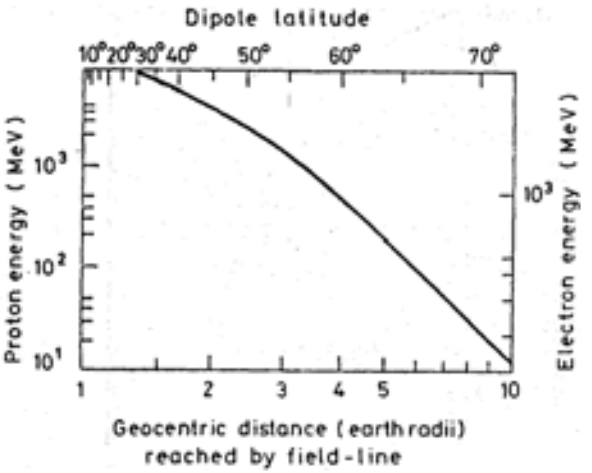


Fig. 20. Penetration of energetic protons into the atmosphere.

D-region ionisation caused by them is usually strongest over the polar caps and consequently the above mentioned radio wave damping. Such short wave absorption events have already been reported in the 1930's well before their true nature was recognized, and were called "polar cap absorption events" or PCAs, a term which is still in use in space weather investigations.

Besides this mainly "technological" consequence of flares there is another very important climatologic one.

Through a complicated chain of chemical reactions the enhanced proton flux causes a strong increase in atmospheric nitrogen which in turn destroys ozone. Therefore a considerable reduction of the total ozone content in the mesosphere and stratosphere has been observed (Fig. 21). Since ozone is a very important climate agens, frequent flares may well contribute to climate effects.

All consequences of the ionisation of energetic solar protons given above in principle apply also to non-solar energetic particles, i.e. galactic cosmic rays (GCR). As explained in previous chapters the flux of GCRs is anti-correlated with solar activity, therefore the ionisation of the mesosphere and stratosphere is generally higher during solar minimum years. This has probably a climatologic impact.

Finally it should be noted that not only the sun can be a cause of space weather effects affecting Earth but also other stars. During cosmic catastrophes, like for instances nova or supernova explosions huge intensities of X- and γ-rays are released. Such an event was registered on 28 August 1998 as consequence of an X-ray burst of a neutron star. The *D*-region experienced a brief spike of ionisation as shown in Fig. 22, despite of the fact that the cause was 23 000 Ly away from Earth. If such an event would occur "close"

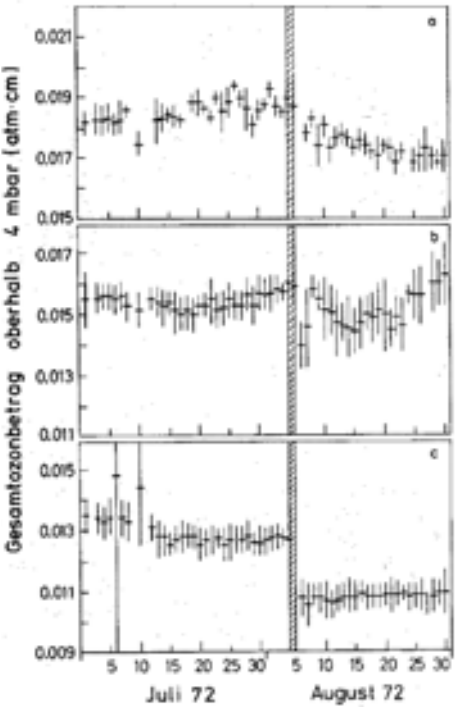


Fig. 21. Total ozone content above 35 km for equatorial (a), mid (b), and high latitudes (c) after the flare of 4. Aug. 1972 (Heath et al., 1977 [8]).

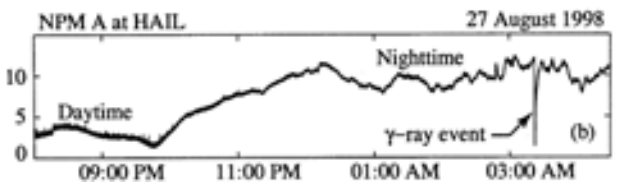


Fig. 22. *D*-region electron density increase (shown here as spike in VLF wave propagation) due to an X-ray burst of the neutron star SGR1900+14.

(e.g. within 50 light years) the terrestrial ozone layer may be destroyed for several years as model calculations show (Ruderman, 1974 [17]). That this would have drastic consequences for the biosphere is obvious!

References

1. Akasofu, S.-I., Space Sci. Rev., **19**, 169, 1970
2. Alcaydé, D., Technical Rpt. 97/53, EISCAT Scienific Assoc., Kiruna, Sweden, 1997

3. Basu, S. and K.M. Groves, in: Space Weather, P. Song, H.J. Singer and G.L. Siscoe (Eds), American Geophys. Union, Washington, D.C. 2001.
4. Chapman, S. and J. Bartels Geomagnetism, Vol. 1 and 2, Oxford, Clarendon Press, 1962
5. Collis, P.N. and M.T. Rietveld, Ann. Geophysicae, **8**, 809-824, 1990.
6. Jakowski, N. A. Wehrenpfennig and S. Heise, Proc. Ionosph. Effects Symp., Alexandria, Virginia, USA, p. 11-18, 2002,
7. Hagfors, T. and K. Schlegel, in: The Century of Space Science, Kluwer Acad. Publ., Dordrecht, 2001
8. Heath, D.F., A.J. Krueger, and P.J. Crutzen, Science, **197**, 886, 1977
9. Kelley, M.C., The Earth's Ionosphere, Acad. Press, San Diego, Ca., 1989
10. Kirkwood, S. and A. Osepian, J. Atmos. Sol. Terr. Phys., **63**, 1907-1922, 2001.
11. Kohl, H., R. Rüster and K. Schlegel (Eds), Modern Ionospheric Science, European Geophysical Society, Katlenburg-Lindau, Germany, 1996
12. Mayaud, P.N., Derivation, meaning and use of geomagnetic indices, American Geophys. Union, Washington D.C., 1980
13. Nevanlinna, H. and E. Kataja, Geophys. Res. Lett., **20**, 2703-2706, 1993.
14. Prölss, G.W., Physik des erdnahen Weltraums, Springer-Verlag, Berlin Heidelberg New York, 2001
15. Rawer, K., Wave propagation in the ionosphere, Kluwer Acad. Publ., Dordrecht, 1993
16. Rees, M.H., Physics and Chemistry of the upper atmosphere, Cambridge Univ. Press, Cambridge, 1989.
17. Ruderman, M.A., Science, **184**, 1079-1081, 1974.
18. Schlegel, K., Vom Regenbogen zum Polarlicht: Leuchterscheinungen in der Atmospäre, Spektrum Akad. Verlag, Heidelberg, 2001
19. Schlegel, K., in: Plasmaphysik im Sonnensystem, K.-H. Glassmeier and M. Scholer, Eds, BI-Wissenschaftsverlag, Mannheim, 1991
20. Schlegel, K., Ann. Geophysicae, **6**, 129-138, 1988.
21. Schlegel, K. and P.N. Collis, J. Atmos. Solar.Terr. Phys., **61**, 217-222, 1999.
22. Walt, M. Introduction to geomagnetically trapped radiation, Cambridge Univ. Press, Cambridge, 1994

Space Weather Effects in the Upper Atmosphere: High Latitudes

Kristian Schlegel

Max Planck Institut für Sonnensystemforschung, 37191 Katlenburg-Lindau, Germany

Abstract. The most important space weather effects on the ionosphere and atmosphere, like the consequences of particle precipitation on conductivities and currents, are described. Following a description of related magnetic signatures on the ground and the relevant geomagnetic indices is a section on the Aurora as a prominent visible aspect of space weather. After a brief discussion of how space weather affects modern communication and navigation, the chapter concludes with an overview of solar flare and cosmic ray related effects.

1 Introduction

This chapter deals with the most important effects of space weather on the ionosphere and atmosphere. It is assumed that the reader is familiar with the basic physics and terminology of both "spheres". For reference in this text a figure is included showing the temperature, ion- and neutral density as a function of altitude (Fig. 1). Recent introductions into modern ionospheric physics are published by Kelley (1989) [9], Kohl et al. (1996) [11], Prölss (2001) [14], and Hagfors and Schlegel (2001) [7]. Many of the space weather effects can be summarized in a flow chart as displayed in Fig. 2. Processes in the magnetosphere, controlled by solar wind and described in previous chapters, cause two major phenomena in the upper atmosphere: particle precipitation and convection of the ionospheric plasma. Particle precipitation enhances the conductivity of the ionospheric plasma, while the plasma convection $\boldsymbol{V}$ in the presence of the Earth's magnetic field causes a system of electric fields due to

$$\boldsymbol{E} = -\boldsymbol{V} \times \boldsymbol{B} \tag{1}$$

Both, the enhanced conductivity σ and the electric field together cause large electric currents mainly in the auroral zone within the dynamo region (90–150 km altitude) according to Ohm's law

$$\boldsymbol{j} = \sigma \boldsymbol{E} \tag{2}$$

Consequences of these currents are Joule heating of the ionospheric plasma which is ultimately transferred to the neutral atmosphere, plasma instabilities, and observable changes of the geomagnetic field on the ground. Further

K. Schlegel, Space Weather Effects in the Upper Atmosphere: High Latitudes, Lect. Notes Phys. **656**, 215–238 (2005)
http://www.springerlink.com/

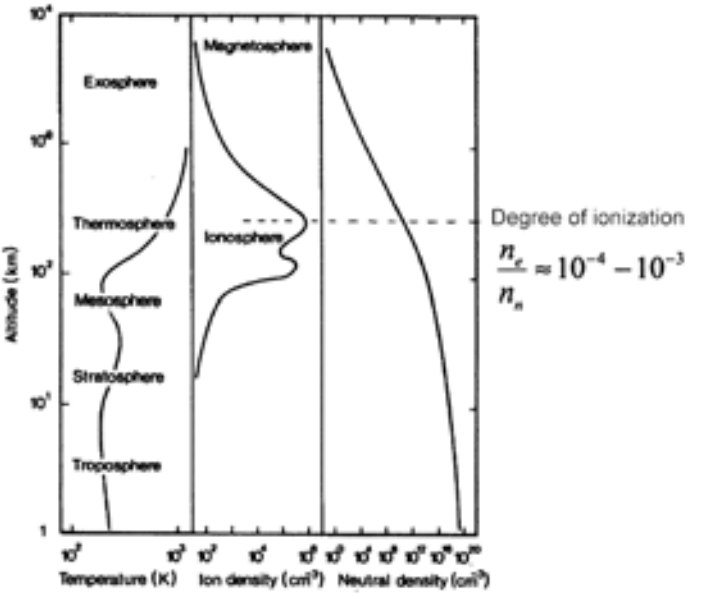


Fig. 1. Basic quantities of the atmosphere and ionosphere as a function of height.

consequences of particle precipitation are another type of heating of the ionospheric plasma and the aurora. The convection of the ionospheric plasma is also partly transferred to the neutral atmosphere by frictional processes and influences the global circulation of the neutral gas (see the chapter by G. Prölss, this volume).

In the following sections we will describe the processes sketched above in more detail.

2 Particle Precipitation

The particles precipitating from the magnetosphere into the ionosphere during space weather events are mainly electrons with energies of a few keV. They are spiralling around the geomagnetic field lines and interact with neutrals and ions by collisions. Neutrals are ionised by these collisions and the ion production as a function of particle energy and altitude can be described with the relation

$$\begin{aligned} q(E_p, z) &= \frac{FE_p}{E_{ion}} \Lambda(s/R) \frac{\rho(z)}{R(E_p)} \\ s &= \int_z^\infty \rho(h) dh \end{aligned} \tag{3}$$

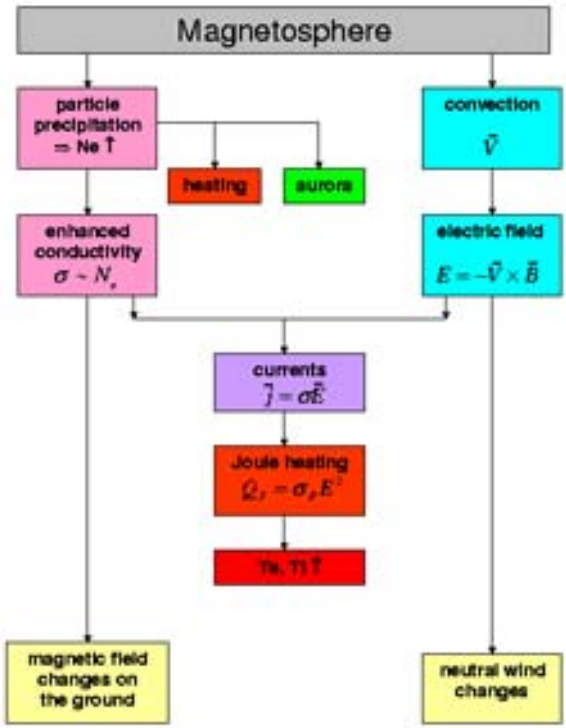


Fig. 2. Flow chart of space weather effects in the ionosphere and atmosphere.

F is the flux of the electrons, R their mean travel distance in air, E_p the peak of the electron energy spectrum, E_{ion} the mean ionisation energy of the neutral gas of 35 eV and $\rho(h)$ the density distribution of the neutral gas. The function Λ describes the energy dissipation of the electrons along their travel path s.

Typical ionisation rates as a function of altitude are plotted in Fig. 3. They are calculated under the assumption of mono-energetic electrons with the energy E_p. They are given for an electron flux of 10^8 particles/s/m^2/sterad and have therefore to be multiplied with the actual flux according to (3). They constitute a good approximation of the real case where the whole energy spectrum of the precipitating electrons has to be taken into account. Such calculations can only be performed in terms of computer simulations (e.g. Kirkwood and Osepian, 2001 [10]).

It is obvious from Fig. 3 that a peak electron energy of a few keV causes a maximum of ionisation in the lower E region. At high latitudes this ionisation can be more than an order of magnitude higher than the "usual" ionisation by solar EUV radiation. The resulting ion density N_i (= electron density N_e, because of charge neutrality) from the ionisation by particles Q_{Part} and solar

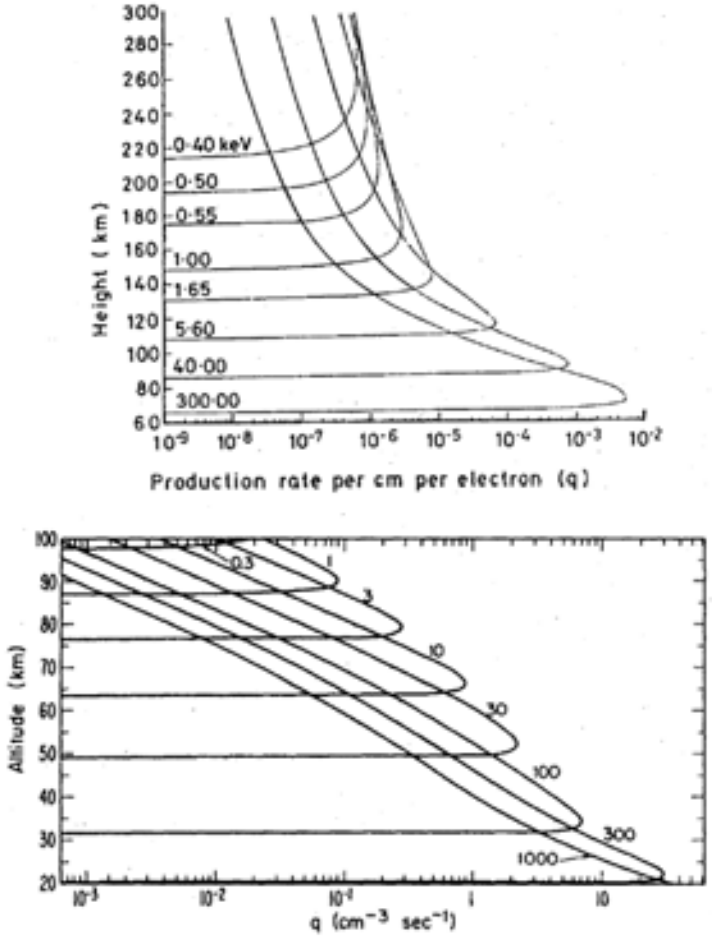


Fig. 3. Ionisation rates caused by precipitation particles (upper panel: electrons, lower panel: protons).

radiation Q_{EUV} can be calculated from the continuity equation

$$\frac{\partial N_i}{\partial t} = Q_{\mathrm{EUV}} + Q_{Part} - L - \mathrm{div}(N_i \boldsymbol{V}_i) \tag{4}$$

where L describes ionisation loss processes and the last term describes transport effects. Figure 4 shows an example of electron densities measured in the high latitude E region with the incoherent scatter technique (Alcaydé, 1997 [2]).

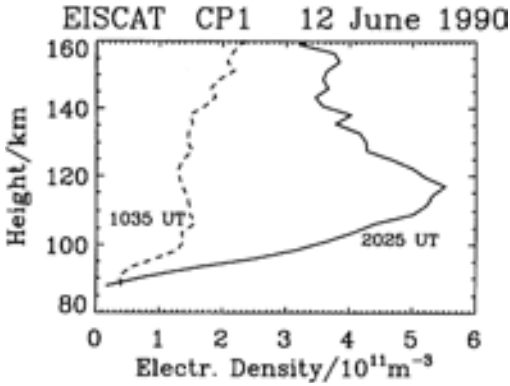


Fig. 4. Example of an electron density enhancement in the auroral E region during particle precipitation (at 20:25 UT) compared to quiet conditions (10:35 UT).

During solar proton events, a consequence of solar flares, high energy protons (up to more than 100 MeV) can precipitate into the ionosphere. Similar equations as (3) describe their ion production rate. Figure 3 also shows production rate profiles for mono-energetic protons. They are able to cause ionisation much deeper in the atmosphere than electrons (see Sect. 7).

3 Conductivities and Currents

Currents in the ionosphere are caused by moving charged particles, thus the current density is given by

$$\boldsymbol{j} = eN_e(\boldsymbol{V}_i - \boldsymbol{V}_e) \tag{5}$$

where e is the elementary charge, and V_i and V_e are ion and electron drifts, respectively. The latter can be calculated from the steady state momentum equations

$$\begin{aligned} &\text{ions}: && e(\boldsymbol{E} + \boldsymbol{V}_i \times \boldsymbol{B}) = m_i \nu_i n(\boldsymbol{V}_i - \boldsymbol{U}) \\ &\text{electrons}: && e(\boldsymbol{E} + \boldsymbol{V}_e \times \boldsymbol{B}) = m_e \nu_e n(\boldsymbol{V}_e - \boldsymbol{U}) \end{aligned} \tag{6}$$

where B is the geomagnetic field (in a coordinate system where $\boldsymbol{B} = (0, 0, B)$), m_i and m_e are ion and electron mass, ν_{in} and ν_{en} the ion-neutral and the electron-neutral collision frequency, respectively, and U the neutral wind.

Combining (5) and (6) yields for the current

$$\begin{aligned}\boldsymbol{j} &= eN_e(\boldsymbol{V}_i - \boldsymbol{V}_e)\\ &= \frac{eN_e}{B}\left\{\left(\frac{\omega_e\nu_{en}}{\omega_e^2+\nu_{en}^2}+\frac{\omega_i\nu_{in}}{\omega_i^2+\nu_{in}^2}\right)(\boldsymbol{E}+\boldsymbol{U}\times\boldsymbol{B})\right.\\ &\quad\left.+\left(\frac{\omega_e^2}{\omega_e^2+\nu_{en}^2}-\frac{\omega_i^2}{\omega_i^2+\nu_{in}^2}\right)(\boldsymbol{E}+\boldsymbol{U}\times\boldsymbol{B})\times\hat{b}\right\}\\ &\quad+e^2N_e\left(\frac{1}{m_e(\nu_{en}+\nu_{ei})}+\frac{1}{m_i\nu_i n}\right)E_\parallel \qquad (7)\\ \omega_{i,e} &= \frac{eB}{m_{i,e}}\end{aligned}$$

The three components of the current density vector are called Pedersen, Hall and parallel current. The Pedersen current flows perpendicular to B and parallel to E, the Hall current perpendicular to B and perpendicular to E, the latter is usually the strongest component. The conductivity can be expressed as a tensor $\tilde{\sigma}$, the current density thus becomes

$$\boldsymbol{j} = \tilde{\sigma}(\boldsymbol{E}+\boldsymbol{U}\times\boldsymbol{B}) \qquad (8)$$

with

$$\tilde{\sigma} = \begin{pmatrix}\sigma_P & \sigma_H & 0\\ -\sigma_H & \sigma_P & 0\\ 0 & 0 & \sigma_\parallel\end{pmatrix} \qquad (9)$$

The three components of this tensor are consequently

$$\begin{aligned}\text{Pedersen conductivity}: \quad \sigma_P &= \frac{eN_e}{B}\left(\frac{\omega_e\nu_{en}}{\omega_e^2+\nu_{en}^2}+\frac{\omega_i\nu_{in}}{\omega_i^2+\nu_{in}^2}\right)\\ \text{Hall conductivity}: \quad \sigma_H &= \frac{eN_e}{B}\left(\frac{\omega_e^2}{\omega_e^2+\nu_{en}^2}-\frac{\omega_i^2}{\omega_i^2+\nu_{in}^2}\right) \qquad (10)\\ \text{parallel conductivity}: \quad \sigma_\parallel &= e^2N_e\left(\frac{1}{m_e\left(\nu_{en}+\nu_{ei}\right)}+\frac{1}{m_i\nu_{in}}\right)\end{aligned}$$

Two important facts can be derived from these equations: (i) all conductivities are proportional to electron density which means that enhanced electron densities as caused by particle precipitation, in turn lead to high conductivities and currents, (ii) all conductivities are strongly height-dependent, because the collision frequencies ν_{in} and ν_{en} are proportional to the neutral density which decreases with altitude (Fig. 1). The latter is usually taken from a neutral atmospheric model (*http://nssdc.gsfc.nasa.gov/space/model/models/msis_n.html*).

Typical quiet-time conductivities are displayed in Fig. 5. Note that σ_H has its peak around 100 km altitude and decreases rapidly with height, whereas

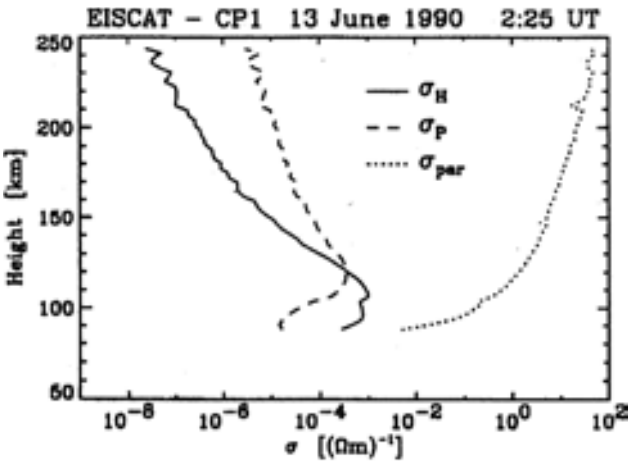


Fig. 5. Typical ionospheric conductivities as derived from EISCAT measurements (Kelley, 1989 [9]).

σ_P has its peak at about 120 km altitude and decreases slowly with height. Although the parallel conductivity is the highest of all three, since the charged particles can move freely along the magnetic field lines, it does not play any role because the parallel electric field is negligibly small. Thus the third term of the current density (7) is not important in the ionosphere.

For practical purposes often the so-called conductance is used. It is the conductivity integrated over the height range where it is most important, i.e.

$$\Sigma_{P,H} = \int_{90\,\mathrm{km}}^{250\,\mathrm{km}} \sigma_{P,H}\,(z)\,dz\ . \tag{11}$$

It has the advantage that it can be conveniently plotted versus time, thus showing the ionospheric variability during space weather events. Figure 6 gives an example which has been computed from measured electron densities (incoherent scatter technique) together with model values of collision frequencies (e.g. Schlegel, 1988 [20]). During the morning and afternoon of that day both, σ_H and σ_P are low, corresponding to quiet conditions where the E region electron density is mainly caused by solar EUV radiation. In the evening and persisting into the next day burst-like conductance enhancements were observed during a geomagnetic storm.

Electric fields in the ionosphere are low during quiet conditions, typically a few mV/m. Generally this field can be regarded as independent of height within the range of interest (90 to several 100 km). During strong geomagnetic storms the field can well exceed 100 mV/m. Figure 7 shows as example the N–S and the E–W components of the electric field during the same time interval as in Fig. 6. Large northward electric fields are observed in the so-

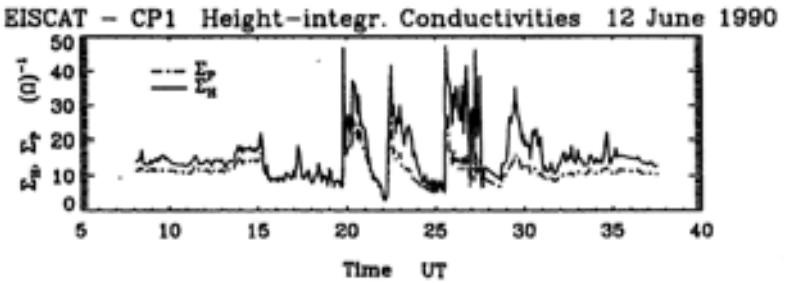


Fig. 6. Hall and Pedersen conductances during a magnetic storm.

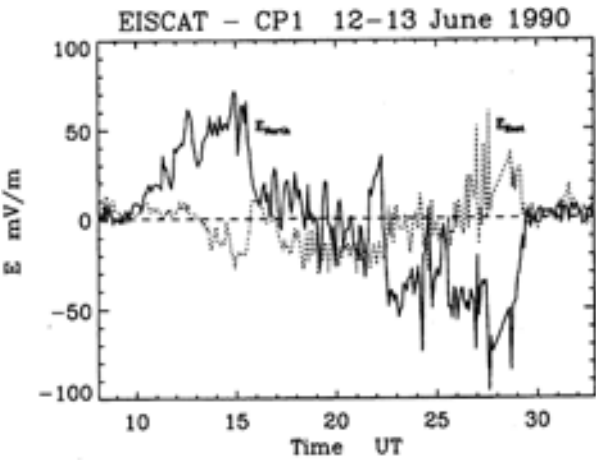


Fig. 7. Typical electric field variations during a magnetic storm.

called afternoon sector, the time between about 11:00 and 17:00 UT. The field turns southward in the subsequent morning sector after about 19:00 UT and remains in this direction until early next morning. This is a very typical electric field behaviour during a magnetic storm. The bursts-like enhancements in the morning sector correspond to substorms.

A typical current density profile during disturbed conditions in the morning sector is plotted in Fig. 8. The peak current flows in a relatively narrow height range between about 90 and 130 km altitude. This current is called the auroral electrojet. It is eastwards directed in the afternoon sector and westwards in the morning sector. The north-south extend of the electrojet is typically about 100 km. Therefore the total current of the electrojet at 1:40 UT on 13 June 1990 was of the order of

$$J = 75\,\mu\mathrm{A/m}^2 \cdot 30000\,\mathrm{m} \cdot 100000\,\mathrm{m} = 0.23 \cdot 10^6\,\mathrm{A}\ . \qquad (12)$$

During very strong magnetic storms it can easily exceed 1 Million A.

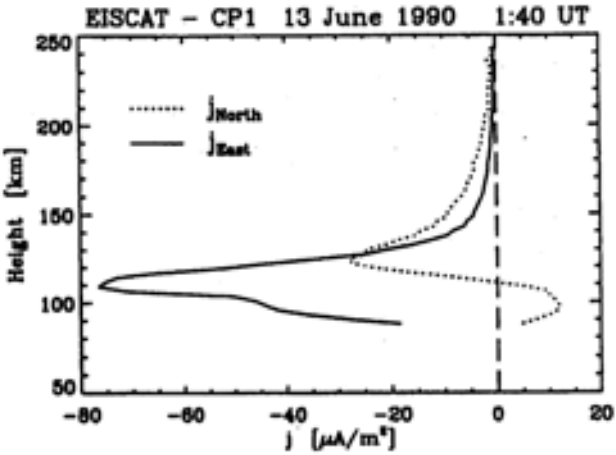


Fig. 8. Current densities as a function of height as derived from EISCAT data during particle precipitation and high electric fields.

The power (per unit volume) dissipated in the E region (Joule heating) by the electrojet is given by

$$P_j = \boldsymbol{j} \cdot \boldsymbol{E} = \sigma_P E^2 \ . \tag{13}$$

The product of the expression for the current (7) and the electric field vector reveals that the Hall current does not contribute to the power, it is a blind current. Microscopically the Joule heating arises from the friction between the charged particles and the neutrals (ν_{in} and ν_{en}). The height-integrated Joule heating

$$Q_J = \Sigma_P \, E^2 \tag{14}$$

can again conveniently be plotted versus time and gives thus an overview of the dissipated energy during a magnetic storm. Figure 9 shows an example which additionally contains the energy input from precipitating particles which can be estimated with the relation

$$Q_P = 1/2 \, \eta \alpha_{eff} N^2 \tag{15}$$

with $\eta = 35\,\mathrm{eV}$ and α_{eff} the effective electron-ion recombination rate. This quantity is generally smaller than the Joule heating, except during brief bursts of precipitation.

Q_J and Q_P constitute the main energy sinks during a magnetic storm, i.e. the energy transferred to the terrestrial atmosphere in a space weather event. It is therefore interesting to compare this energy with the total energy transferred to the magnetosphere by the solar wind. During the geomagnetic storm of 10 January 1997 the energy dissipated by Joule heating over the

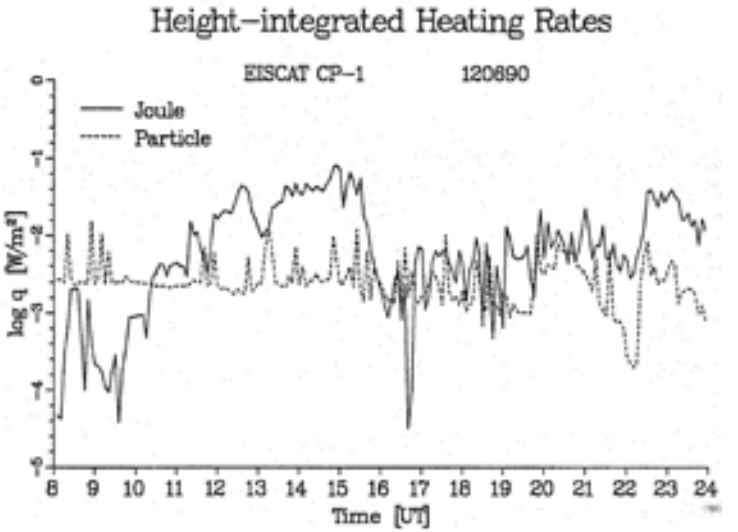


Fig. 9. Joule and particle heating in the auroral ionosphere during a magnetic storm.

whole auroral zone was estimated to 13000 TJ which is about 40% of the average solar wind energy input into the magnetosphere (Schlegel and Collis, 1999 [21])).

The energy transferred to the atmosphere causes first of all the ion and electron gas to be heated, but ultimately this energy is passed to the neutral gas. It causes a considerable expansion of the auroral atmosphere. This has important consequences for satellites orbiting the Earth at altitudes below about 500 km as already mentioned in the previous chapter by G. Prölss.

It should be noted in this context that a heating of the terrestrial atmosphere occurs regularly within the solar cycle, apart from magnetic storms. Whereas the visible and infrared part of the solar spectrum does only marginally change during the solar cycle, the EUV-flux in the wavelength range below 100 nm is increased by more than a factor of three during solar activity maximum years. This yields a higher energy input into the upper atmosphere, since this part of the solar radiation is mainly absorbed at altitudes above 100 km. Consequently not only the neutral gas density and temperature but also the ionisation is increased, as demonstrated in Fig. 10. Apart from the consequences for satellite trajectories this also leads to important differences in short wave radio propagation during the solar cycle, as known for more that 70 years. Even radio amateurs enjoy the greater distances to be covered during solar maximum years.

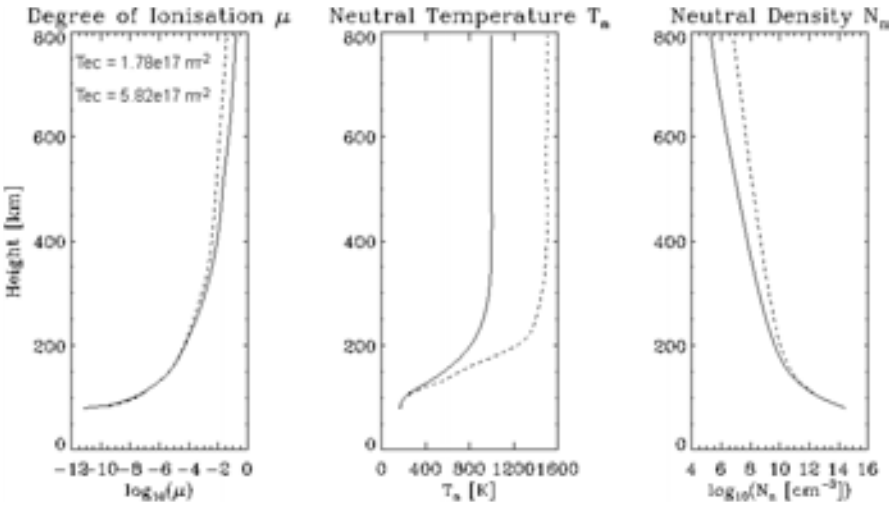


Fig. 10. Difference between high and low solar activity in atmospheric ionisation, temperature, and density.

4 Magnetic Signatures on the Ground and Geomagnetic Indices

The auroral electrojet causes distinct perturbations of the geomagnetic field which can be monitored with magnetometers on the ground. Although the Hall current is a blind current, it usually causes the strongest variations $\Delta \boldsymbol{B}$. According to the "right-hand-rule" the magnetic perturbation appears mainly in the N–component of $\Delta \boldsymbol{B}$ in the evening sector and in the S–component in the morning sector. With the help of N–S aligned magnetometer chains the location and extend of the electrojet can be well established, Fig. 11 shows an example.

Equally important are ground-based magnetometers for the derivation of geomagnetic indices which are widely used to characterize space weather events in a quantitative manner. Since the pioneering work of the German geophysicist Julius Bartels (1899–1964) geomagnetic storms are characterised by the index Kp (Chapman and Bartels, 1962 [4]). Bartels who introduced this index in 1949 derived it from the largest variation of the horizontal magnetic field component during a 3–hour interval from a single magnetometer station, using a quasi-logarithmic scale. This so-called K index was then averaged over 13 globally distributed stations, applying special weighting functions, in order to obtain the Kp index where p stands for "planetary". Kp runs from 0 (very quiet) to 9 (very disturbed) and is further subdivided using the subscripts $-, 0, +$ (e.g. $1-, 1_0, 1+, 2-, ...$), this yields 28 steps in total. Bartels also developed a convenient representation of Kp in terms of musical notes (Fig. 12).

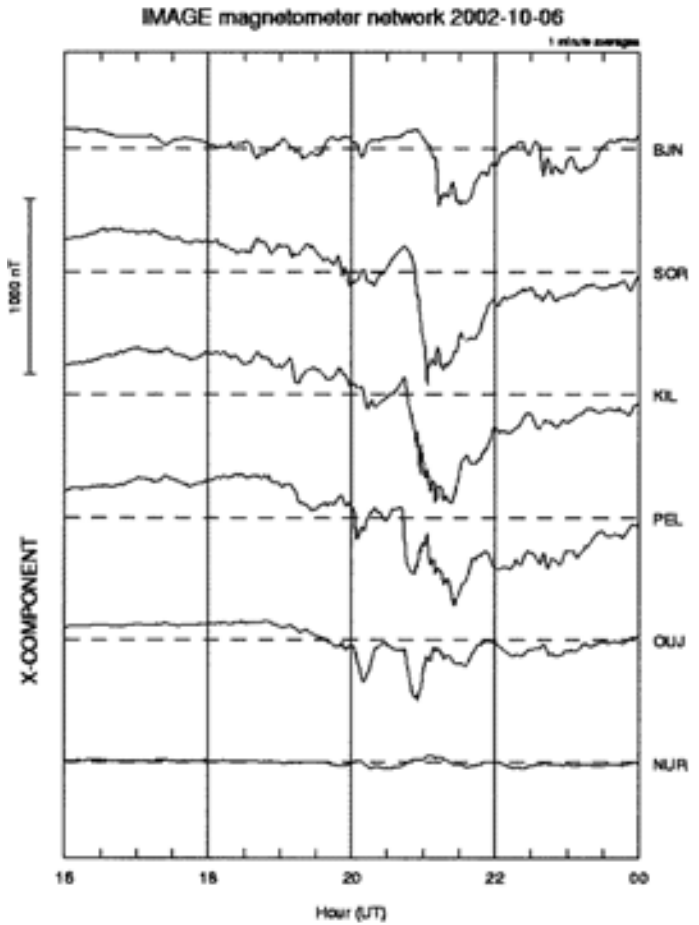


Fig. 11. Record of the x-component of the geomagnetic field from a part of the IMAGE magnetometer network. The instruments are apprximately meridionally aligned from southern Finland (NUR) to Bear Island (BJN).

As already mentioned, Kp is expressed in a logarithmic scale and consequently not very well suited for averaging. Bartels therefore introduced the linear equivalent ap where $\mathrm{Kp} = 9_0$ corresponds to $\mathrm{ap} = 400\,\mathrm{nT}$. The Ap index is a mean over eight 3–hour intervals of ap, i.e. over a full day. It consequently characterises not only the strength but also the duration of the strongest phase of a storm.

Bartels was able to derive both indices back to 1932, for earlier years not enough magnetometer stations were available. For many years the University of Göttingen issued the Kp and Ap indices, but since beginning of 1997 this task has been taken over by the Adolf-Schmidt Observatorium für Ge-

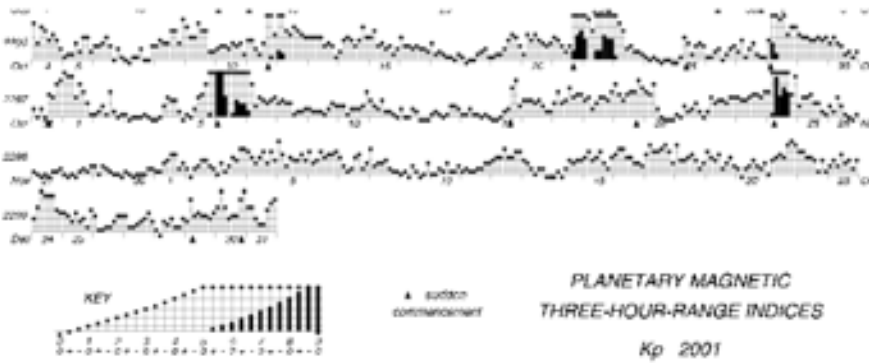


Fig. 12. Bartels' Kp-notation as musical notes. Every line represents 28 days, the average solar rotation. Recurrency trends can therefore be easily detected.

omagnetismus in Niemegk/Germany (*http://www.gfz-potsdam.de/pb2/pb23/GeoMag/niemegk/obs_eng.html*).

In order to characterize geomagnetic storms before 1932 a different index, the so-called AA index was developed. It is similar as Ap but is derived from the magnetograms of only two stations, one on the northern (England) and one on the southern hemisphere (Australia). Since both stations have recorded the geomagnetic field since 1868, it was possible to derive aa (3–h interval) and AA (full day) back to this year. Finnish scientist have recently pushed the AA-records even further back (Nevanlinna and Kataja, 1993 [13]).

The so far mentioned indices characterise geomagnetic variations particularly at high and midlatitudes which are mainly related with the auroral electrojet. The magnetic variations due to the ring current (see Chapter by G. Prölss) are described by the Dst index. It is derived since 1957 from the horizontal magnetic field component measured at 4 stations near the equator. The magnetic field of the ring current is directed opposite to the main geomagnetic field, consequently strong disturbances are characterised by large negative Dst excursions.

Finally, magnetic disturbances at very high latitudes are characterised by the AE index which is derived from magnetic records of 12 stations at auroral latitudes. Details of the derivation of all indices can be found in Mayaud (1080) [12], their values are accessible through the internet (*http://www.cetp.ipsl.fr/~isgi/homepag1.htm, http://spidr.ngdc.noaa.gov/spidr/html*). A map with the stations for the various indices is given in Fig. 13. The convenience of geomagnetic indices is demonstrated with Table 1, listing the 10 strongest storms of the past century.

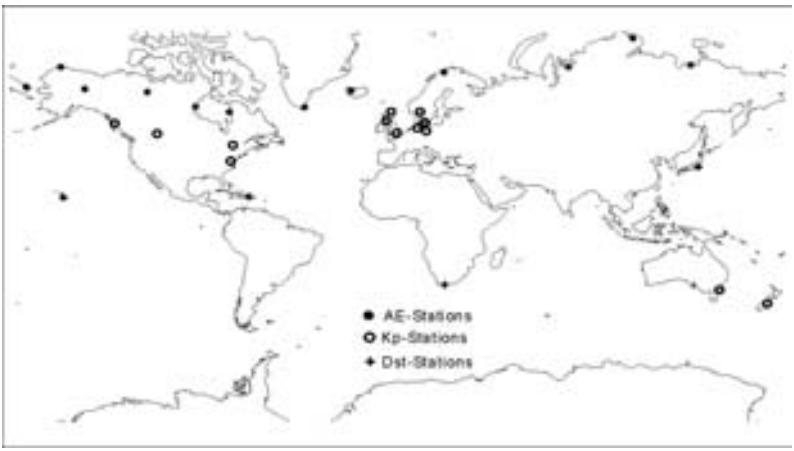


Fig. 13. Map of the location of magnetometer stations from which the various magnetic indices are derived.

Table 1. The 10 strongest geomagnetic storms in the past century in descending order. The second column gives AA*= Σaa of the most strongest 24–h interval (AA, without the star corresponds to one day.) The third column gives the maximal Kp (since 1932) and the fourth the minimal Dst (since1957) in the corresponding interval. The last column shows the location of auroral observations most near to the equator during the storms. Si refers to a list of auroral observations compiled by Silverman ranging from 686 BC to 1951 AD, Sch to W. Schröder (private communication), A to other sources.

Date	AA* [nT]	max. Kp	min. Dst [nT]	Auroral Observation nearest to the Equator (geogr. Latitude)
1989 13./14. March	441	9_0	–589	A: Florida Keys, ($\Phi \approx 24^\circ$ N)
1941 18./19. Sept.	429	9–	–	Si: Florida, ($\Phi \approx 29^\circ$ N)
1940 24./25. March	377	9_0	–	Si: Korfu, ($\Phi = 39^\circ$ N)
1960 12./13. Nov.	372	9_0	–339	A: Atlantic, ($\Phi = 28^\circ$ N)
1959 15./16. July	357	9_0	–429	Sch: $\Phi \approx 48^\circ$ N
1921 14./15. May	356	–	–	Si: Samoa, ($\Phi = 14^\circ$ S)
1909 25./26. Sept.	333	–	–	Si: Mallorca, ($\Phi = 39^\circ$ N)
1946 28./29. March	329	9_0	–	Si: Queensland, ($\Phi \approx 27^\circ$ S)
1928 7./8. July	325	–	–	Si: Atlantic, ($\Phi = 24^\circ$ N)
1903 31.10./1.11.	324	–	–	Si: Bamberg, ($\Phi = 50^\circ$ N)

5 Aurora

Aurora is the only visible and pleasant aspect of space weather. They are caused by the aforementioned keV-particles precipitating from the magnetosphere into the upper atmosphere. At altitudes between about 500 and 90 km these particles interact with atmospheric constituents, mainly N_2, O_2 and O. These constituents are excited and subsequently radiate the excitation energy

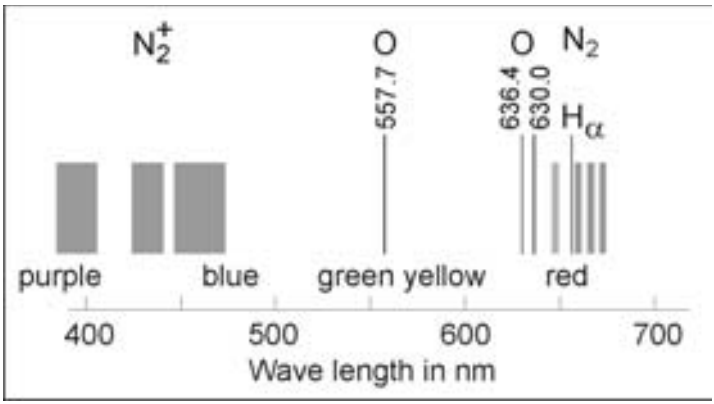


Fig. 14. Simplified spectrum of auroral emissions in the visible range.

over a broad spectrum (infrared, visible, ultraviolet). It should be noted that only a small part of the emissions are caused by direct collisional excitation through the precipitating particles or their secondaries, the major part is released in chemical reactions which are in turn induced or affected by these particles. Figure 14 shows a simplified spectrum of auroral emissions in the visible range.

The predominant green colour of aurora is caused by the following reactions

$$\mathrm{N}^+ + \mathrm{OH}^2 \quad \rightarrow \quad \mathrm{NOH}^+ + \mathrm{OH}\ \left(^1\mathrm{S}\right),$$

the excited oxygen atom then transits in a lower excitation state by emitting a photon

$$\mathrm{O}\ \left(^1\mathrm{S}\right) \quad \rightarrow \quad \mathrm{O}\ \left(^1\mathrm{D}\right) + \mathrm{Photon}\,(555.7\,\mathrm{nm}).$$

Red light is emitted when this metastable state goes to the ground state

$$\mathrm{O}\ \left(^1\mathrm{D}\right) \quad \rightarrow \quad \mathrm{O}\ \left(^3\mathrm{P}\right) + \mathrm{Photon}\,(630\,\mathrm{nm}).$$

The O $\left(^1\mathrm{D}\right)$-state can also directly be excited by dissociative recombination:

$$\mathrm{O}_2^+ + e \quad \rightarrow \quad \mathrm{O} + \mathrm{O}\ \left(^1\mathrm{D}\right)$$

Nitrogen molecules can be ionised and excited by primary electrons:

$$\mathrm{N}_2 + \mathrm{e}_{\mathrm{primary}} \quad \rightarrow \quad \mathrm{N}_2^+\left(\mathrm{B}^2\Sigma_u^+\right) + \mathrm{e}_{\mathrm{primary}} + \mathrm{e}_{\mathrm{sekundary}}$$

This excited state of the ionised nitrogen molecule is a vibrational state. In the transition to other vibration states a whole band of colours in the blue-violet range is emitted. Emissions of the neutral nitrogen molecule fall within the red and ultraviolet bands (mode details of emissions, see Rees, 1998 [16]).

The green and the red line (the latter is actual a doublet) are so-called forbidden lines. The corresponding excitation states have a relatively long

Table 2.

IBC	Intensity of the 557.7 nm line (kR = kilo Rayleigh)	Comparable brightness
I	1 kR	as the milky way
II	10 kR	moonlight on thin cirrus
III	100 kR	moonlight on cumulus
IV	1000 kR	full moon

life time of 1 s (green) and 110 s (red). Under normal pressure at ground level these excited states would be immediately quenched by collisions with other atmospheric constituents. Only at altitudes above 100 km and pressures below 0.1 Pa, the mean time between two collisions is longer than the excitation life time and the de-excitation by emission becomes possible. The association of these lines to atomic oxygen was therefore a longstanding problem to spectroscopists and was finally solved not before 1932.

The brightness of aurora is characterized by the "international brightness coefficient" (IBC) according to four classes (1 Rayleigh = 106 photons/cm^2/s/sterad) as listed in Table 2.

The special topology of the geomagnetic field lines extending into the magnetospheric tail cause the aurora to be confined mainly to a ring around the magnetic poles, the so-called auroral oval (Fig. 15). Within this ring which is located at about 70° geomagnetic latitude and has a typical width of several 100 km during not too disturbed conditions, aurora occurs most frequently. During very strong space weather events the auroral oval expands towards the equator and can easily reach mid latitudes. Due to the smaller dip angle of the field lines the auroral particles experience a longer travel time through the atmosphere and therefore aurora appears mainly at altitudes above 200 km as a red glow. These red colours were associated with blood by our ancestors, and therefore aurora was regarded as a bad omen for war and diseases (Schlegel, 2001 [18]). The forms of aurora depend on the topology of the currents flowing from the magnetotail into the polar regions. In principle two basic manifestations exist: diffuse aurora (unstructured, extended) and discrete aurora (arcs, veils, bands, localised). Some common forms are sketched in Fig. 16. The diffuse aurora is caused by particles in the 100 eV range which are scattered into the loss cone, and appear mainly at altitudes above 150 km. The energy of particles causing the discrete aurora on the other hand, is of the order of several keV as already mentioned, considerably lower than the mean energy of the particles in the plasma sheet of the tail, their origin. The particles have therefore to be accelerated along the field lines. The nature of this acceleration is still under debate, several different mechanisms are discussed (e.g. Schlegel, 1991 [19]).

Auroral particles causing the so far mentioned aurora are electrons. The "proton aurora" is much more rare and is caused by energetic protons which

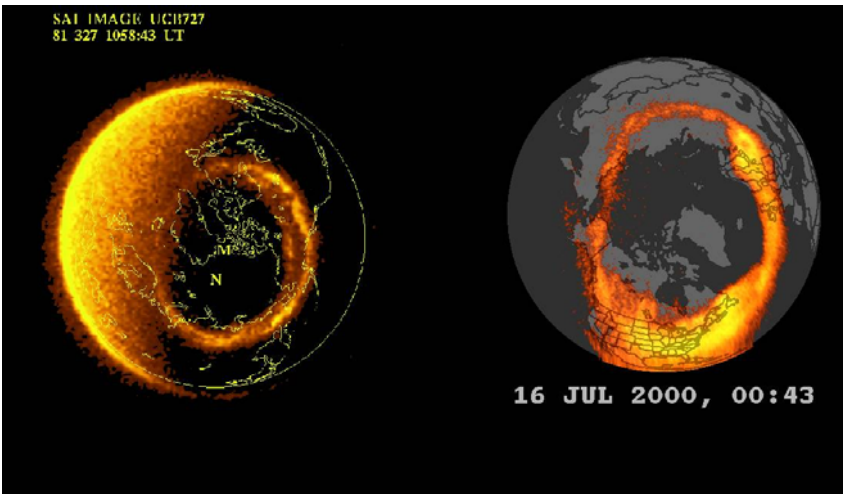


Fig. 15. Auroral oval during quiet (left) and disturbed (right) conditions.

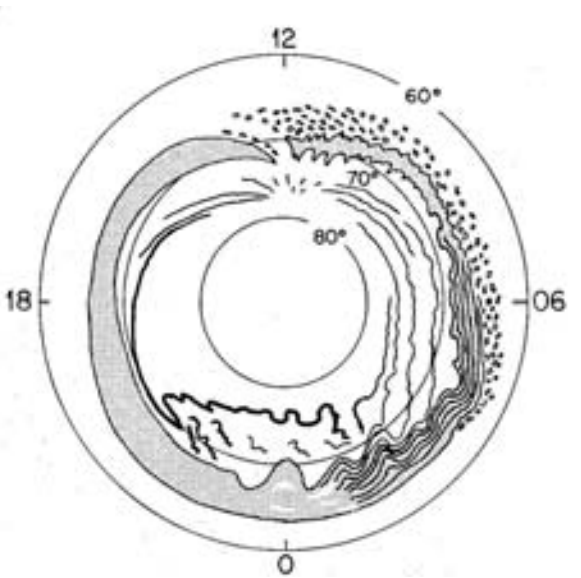


Fig. 16. Schematic representation of the main auroral forms as a function of local time for latitudes $> 60°$. The shaded area characterises diffuse aurora, the thick line a quiet arc which transforms into folded bands after about 21:00 LT. In the morning hours patchy aurora can often be found at the southern rim of the auroral oval. The short thin lines around local noon at about 75° latitude are daylight aurora (Akasofu, 1970 [1]).

Fig. 17. Four auroral displays with different colours and forms. The lower right one was photographed at mid-latitudes (near Düsseldorf).

are decelerated in the atmosphere by collisions and finally transformed to excited neutral hydrogen by charge transfer:

$$M \ + \ \mathrm{H} \quad \rightarrow \quad M^{+} \ + \ \mathrm{H}^{*},$$

where M is any neutral constituent. The excited hydrogen atoms emit L_α (121.57 nm, UV) or H_α (656.3 nm, red). The latter cannot be distinguished by eye from the red oxygen light (Fig. 14). Proton aurora is generally diffuse and often associated with PCA events (Sect. 7).

There are plenty of internet pages with splendid auroral photos, e.g. *http://www.meteoros.de, http://www.exploratorium.edu/learning_studio/auroras/, http://sgo.fi/Pictures/, http://www.pi.physics.uiowa.edu/vis/*. A brief collection is printed in Fig. 17.

6 Consequences of Electron Density Enhancements and Fluctuations

The enhancement of electron density by precipitating particles as described in Sect. 2 has important consequences on communication and navigation.

Although the importance of HF communication which is most strongly affected, has decreased in recent years, it still plays a role in many countries. It is therefore necessary to forecast possible changes of HF propagation during space weather events.

The propagation of electromagnetic waves in the ionosphere is described by the magneto-ionic theory (Rawer, 1993 [15]). One important equation is the index of refraction of the waves which in its simplest form reads

$$n^2 = 1 - \frac{\omega_P^2}{\omega^2} \tag{17}$$

$$\text{with the plasma frequency } \omega_P = \sqrt{\frac{N_e e^2}{\epsilon_0 m_e}} \ .$$

It is obvious from this equation that the propagation of a wave with frequency ω depends strongly on the plasma frequency and thus on the electron density. Waves used for communication under quiet conditions may not reach their destination (for instance, when n becomes imaginary) under conditions with enhanced electron density.

Eq uation (16) indicates a strong decrease of ionospheric propagation effects for large frequencies, for $\omega \gg \omega_P$, the refractive index approaches unity which means propagation in vacuum space. But even for GHz radio waves the propagation effects are not negligible in certain cases, for instance in GPS navigation.

A very important ionospheric quantity in this context is the total electron content

$$TEC = \int_P N_e \ ds \ , \tag{18}$$

where the integral is taken over the signal path from the ground station to the satellite. Typical values are of the order of 50–150 TECU (TEC-units, 10^{16} electrons/cm^2). Due to enhanced electron densities during particle precipitation changes of the order of several 10 TECU can easily occur. Figure 18 shows an example. In case of GPS measured distances the TEC change translates into errors of

$$D(\text{mm}) = 2.16\, TEC \text{ (in TECU)} \tag{19}$$

It should be noted that electron density enhancements may not only occur at high latitudes, but can also be convected towards lower latitudes as so called patches.

The strong currents in the auroral E-region cause plasma instabilities which lead to a structuring of the normally uniform plasma. The formed plasma irregularities have a broad range of scale lengths λ_{irr}, from kilometres

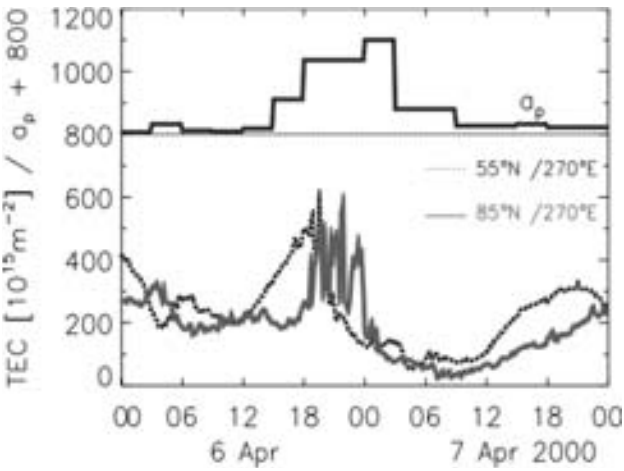


Fig. 18. Variation of TEC during a magnetic storm, the upper panel shows ap (Jakowski et al., 2002 [6]).

to tenth of meters and can therefore cause constructive interference with radio waves. This can lead either to strong backscatter or to forward scatter of radio waves in a wide frequency band whenever

$$\lambda_{radio\ wave} = 2\lambda_{irr} \tag{20}$$

A corresponding effect which is often observed during space weather events is the overrange of VHf signals, e.g. that taxi drivers in Hamburg can listen to their colleagues in Helsinki over their usual communication channels. Radio amateurs too use this "auroral scatter" as they termed it, for long range communication.

Even satellite signals in the GHz range are affected in such cases. This "radio scintillation" causes amplitude and phase fluctuations of satellite signals and thereby disturbs the communication and also degrades the accuracy of GPS measurements (Basu and Groves, 2001 [3]).

7 Solar Flare and Cosmic Ray Related Effects

As mentioned in previous chapters during and after a solar flare the flux of high energy protons as well as of X-rays is enhanced at the Earth by several orders of magnitude. During the very strong flare on 18 August 1979 for instance, the X-ray flux in the wavelength range 0.029–0.048 nm increased by a factor of 2000, and that of the 0.05–0.8 nm by a factor of 280. Solar X-rays play in general an important role in the ionisation of the ionospheric D-region. An enhancement of their flux can therefore considerably increase the electron density in the height range 80–100 km (Collis and Rietveld, 1990

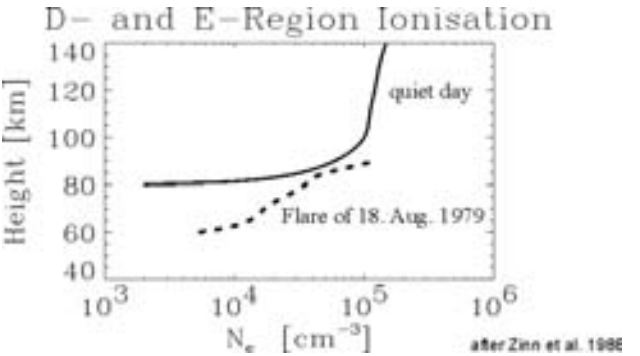


Fig. 19. Electron density during quiet conditions and during the solar flare of 18 Aug. 1979.

[5]). A similar ionisation increase cause the high energy protons which can penetrate well down into the stratosphere (see Fig. 3 above). Figure 19 shows an example of the *D*-region electron density increase during the above mentioned flare. Whereas the X-rays reach the Earth only about 8 min after the flare onset, the energetic protons need a travel time of the order of one hour. The X-ray flux increase is peak-like with a duration of only about 10 min, whereas the enhanced proton flux usually pertains for several days. Thus the large electron densities in the mesosphere and stratosphere maintain for a similar time. High electron densities together with the high electron-neutral collision frequencies at *D*-region heights cause a strong damping of electromagnetic waves according to magneto-ionic theory (Rawer, 1993 [15]). Thus short (MHz) and medium (kHz) wave communication is strongly affected in such cases.

In the vicinity of the Earth the energetic protons gyrate around the geomagnetic field lines according to Störmer's theory (e.g. Walt, 1994 [22]). An important quantity for their propagation is the "magnetic rigidity"

$$R = \frac{pc}{Ze} \ , \tag{21}$$

where p is the particle momentum, c the velocity of light and Z their charge number (this formula also applies to particle with $Z > 1$, e.g. alpha particles). All particles with the same rigidity have the same orbit parameters. It can be shown that all particles with a critical rigidity

$$R_c = 14.9 \cos^4 \lambda_c \tag{22}$$

reach geomagnetic latitudes $\lambda \geq \lambda_c$, or differently expressed, all particles with $R \geq R_c$ can reach the geomagnetic latitude λ_c. This is explained in Fig. 20: protons with energies $E_p < 100$ MeV will penetrate the Earth's atmosphere only at high latitudes; the higher their energy, the more lower latitudes they can reach. Since the peak of solar protons is normally below 100 MeV, the

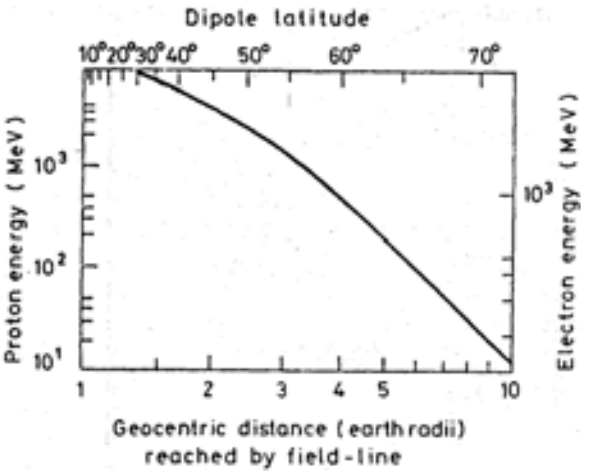


Fig. 20. Penetration of energetic protons into the atmosphere.

D-region ionisation caused by them is usually strongest over the polar caps and consequently the above mentioned radio wave damping. Such short wave absorption events have already been reported in the 1930's well before their true nature was recognized, and were called "polar cap absorption events" or PCAs, a term which is still in use in space weather investigations.

Besides this mainly "technological" consequence of flares there is another very important climatologic one.

Through a complicated chain of chemical reactions the enhanced proton flux causes a strong increase in atmospheric nitrogen which in turn destroys ozone. Therefore a considerable reduction of the total ozone content in the mesosphere and stratosphere has been observed (Fig. 21). Since ozone is a very important climate agens, frequent flares may well contribute to climate effects.

All consequences of the ionisation of energetic solar protons given above in principle apply also to non-solar energetic particles, i.e. galactic cosmic rays (GCR). As explained in previous chapters the flux of GCRs is anti-correlated with solar activity, therefore the ionisation of the mesosphere and stratosphere is generally higher during solar minimum years. This has probably a climatologic impact.

Finally it should be noted that not only the sun can be a cause of space weather effects affecting Earth but also other stars. During cosmic catastrophes, like for instances nova or supernova explosions huge intensities of X- and γ-rays are released. Such an event was registered on 28 August 1998 as consequence of an X-ray burst of a neutron star. The *D*-region experienced a brief spike of ionisation as shown in Fig. 22, despite of the fact that the cause was 23 000 Ly away from Earth. If such an event would occur "close"

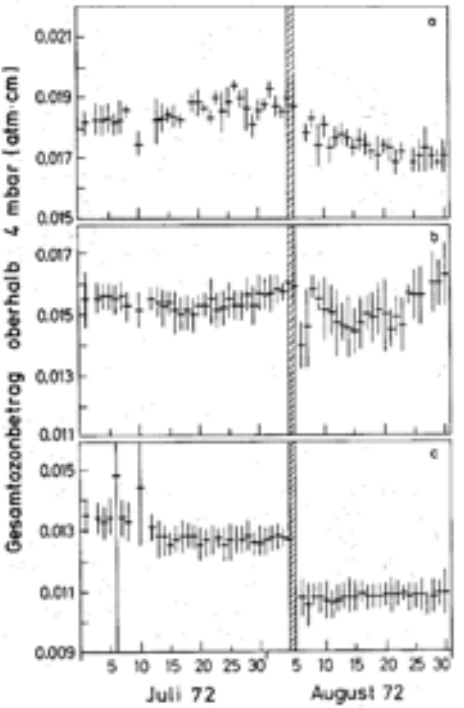


Fig. 21. Total ozone content above 35 km for equatorial (a), mid (b), and high latitudes (c) after the flare of 4. Aug. 1972 (Heath et al., 1977 [8]).

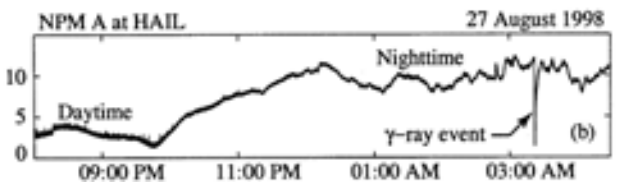


Fig. 22. *D*-region electron density increase (shown here as spike in VLF wave propagation) due to an X-ray burst of the neutron star SGR1900+14.

(e.g. within 50 light years) the terrestrial ozone layer may be destroyed for several years as model calculations show (Ruderman, 1974 [17]). That this would have drastic consequences for the biosphere is obvious!

References

1. Akasofu, S.-I., Space Sci. Rev., **19**, 169, 1970
2. Alcaydé, D., Technical Rpt. 97/53, EISCAT Scienific Assoc., Kiruna, Sweden, 1997

3. Basu, S. and K.M. Groves, in: Space Weather, P. Song, H.J. Singer and G.L. Siscoe (Eds), American Geophys. Union, Washington, D.C. 2001.
4. Chapman, S. and J. Bartels Geomagnetism, Vol. 1 and 2, Oxford, Clarendon Press, 1962
5. Collis, P.N. and M.T. Rietveld, Ann. Geophysicae, **8**, 809-824, 1990.
6. Jakowski, N. A. Wehrenpfennig and S. Heise, Proc. Ionosph. Effects Symp., Alexandria, Virginia, USA, p. 11-18, 2002,
7. Hagfors, T. and K. Schlegel, in: The Century of Space Science, Kluwer Acad. Publ., Dordrecht, 2001
8. Heath, D.F., A.J. Krueger, and P.J. Crutzen, Science, **197**, 886, 1977
9. Kelley, M.C., The Earth's Ionosphere, Acad. Press, San Diego, Ca., 1989
10. Kirkwood, S. and A. Osepian, J. Atmos. Sol. Terr. Phys., **63**, 1907-1922, 2001.
11. Kohl, H., R. Rüster and K. Schlegel (Eds), Modern Ionospheric Science, European Geophysical Society, Katlenburg-Lindau, Germany, 1996
12. Mayaud, P.N., Derivation, meaning and use of geomagnetic indices, American Geophys. Union, Washington D.C., 1980
13. Nevanlinna, H. and E. Kataja, Geophys. Res. Lett., **20**, 2703-2706, 1993.
14. Prölss, G.W., Physik des erdnahen Weltraums, Springer-Verlag, Berlin Heidelberg New York, 2001
15. Rawer, K., Wave propagation in the ionosphere, Kluwer Acad. Publ., Dordrecht, 1993
16. Rees, M.H., Physics and Chemistry of the upper atmosphere, Cambridge Univ. Press, Cambridge, 1989.
17. Ruderman, M.A., Science, **184**, 1079-1081, 1974.
18. Schlegel, K., Vom Regenbogen zum Polarlicht: Leuchterscheinungen in der Atmospäre, Spektrum Akad. Verlag, Heidelberg, 2001
19. Schlegel, K., in: Plasmaphysik im Sonnensystem, K.-H. Glassmeier and M. Scholer, Eds, BI-Wissenschaftsverlag, Mannheim, 1991
20. Schlegel, K., Ann. Geophysicae, **6**, 129-138, 1988.
21. Schlegel, K. and P.N. Collis, J. Atmos. Solar.Terr. Phys., **61**, 217-222, 1999.
22. Walt, M. Introduction to geomagnetically trapped radiation, Cambridge Univ. Press, Cambridge, 1994

Space Weather Effects on Technology

Eino Valtonen

Space Research Laboratory, Department of Physics, University of Turku,
20014 Turku University, Finland

Abstract. Space weather effects on technology are discussed. After a brief general overview of the consequences of space weather on various technologies and a summary of the space environment, the effects of plasmas and particle radiation on systems in space are considered in more detail. The processes responsible for spacecraft surface charging and internal charging and the consequences of subsequent discharges are examined. The concepts of total ionising dose, displacement damage, and single event effects are introduced. The principles how they affect semiconductor devices are discussed and the basic ideas for calculating these quantities are presented. Finally, plasma and particle-induced interference and background in scientific sensors are considered. Examples of various types of space weather effects on space systems are given.

1 Introduction

With increasing complexity and susceptibility of technological systems to perturbations in conditions in near-Earth space, and with growing dependence of functionalities of society on such systems, space weather has gained considerable importance in the discipline of solar-terrestrial research. In recent years, this has led to implementation or preparation of large-scale national (NSWP, 2000 [1]) and international (Hapgood, 2001 [2]) space weather programs.

Space weather effects on technology are manifold and can be experienced in deep space as well as on the surface of the Earth. Space weather effects are linked to the conditions in space environment, which largely are controlled by the Sun. The ultimate source of most of the space weather effects is the Sun, although many of them are caused by various secondary processes. The topic of this chapter is limited to space weather effects on spacecraft systems with some emphasis on technologies employed in scientific payloads. In this regard, space weather can be primarily described in terms of plasmas and time and energy variable fluxes of particles.

In Sect. 2, in order not to totally overlook the effects experienced inside the atmosphere of the Earth, a general overview of various types of space weather effects on technology is given. Section 3 summarises the most important characteristics of space environment. The main topics, plasma and radiation effects on systems in space and the physical mechanisms responsible for these effects, are discussed in Sects. 4 and 5, respectively. In Sect. 6, a summary and final remarks are presented.

E. Valtonen, Space Weather Effects on Technology, Lect. Notes Phys. **656**, 241–273 (2005)
http://www.springerlink.com/

2 Overview of Space Weather Effects on Technology

Space weather may have impact on technological systems in space, on ground, as well as on systems relying in their operation on the conditions in the atmosphere. Illustrative examples are given in Odenwald (2001) [3]. A brief historical review of phenomena, which today are known as space weather effects, and an overview of the current technologies that can be affected by solar-terrestrial processes have been recently presented in Lanzerotti (2001) [4]. Various space weather phenomena and effects were analysed in Koskinen et al. (2001) [5], and a set of excellent domain, phenomena, and system oriented catalogues were compiled.

The earliest observed space weather effects were anomalous currents in telegraph systems (Lanzerotti, 2001 [4]), now known to be caused by geomagnetically induced currents. These currents are driven by electric fields induced in the Earth and in conductors at or near the surface of the Earth by variations in the strength and direction of the geomagnetic field. Such magnetic field variations result from greatly increased electrical current systems in the magnetosphere and the ionosphere during geomagnetic storms. In present day, the most serious consequences of geomagnetically induced currents can be seen in power transmission systems, long telecommunications cables, and pipelines (Boteler et al., 1998 [6]; Pirjola et al., 2000 [7]; Molinski et al., 2000 [8]; Pulkkinen et al., 2001 [9]).

The quality of high-frequency wireless communications has long been known to be dependent on the conditions in the ionosphere. Changes in absorption and reflection produced by solar activity can significantly alter the propagation of radio signals from a location to another on the Earth's surface. Earth-to-satellite communication links also suffer from ionospheric disturbances (Basu et al., 2002 [10]). Turbulences and irregularities in the ionosphere scatter radio waves and cause temporal fluctuations in intensity and phase, called scintillations. Such plasma processes in the ionosphere can introduce positioning errors in the (Skone, 2001 [11]).

Space storm effects on satellites have been recently discussed in Baker (2001) [12]. In Panasyuk (2001) [13] radiation hazards to space missions were considered, and space weather effects on operations in space were discussed in Shea and Smart (1998) [14]. The results from CRRES, a dedicated space environment mission, have been analysed in Brautigam (2002) [15], summarising how CRRES has changed our understanding of magnetospheric radiation hazards. Some recent anomalies in communications satellites were analysed in Gubby and Evans (2002) [16] and guidelines for minimising susceptibility of satellites to space weather were presented. The most frequently encountered problems in space are various plasma and radiation effects. These will be examined in detail in Sects. 4 and 5 of this chapter. Other types of effects of concern are drag and attitude perturbations by the expanding atmosphere due to bursts of solar X- and UV-radiation, material and surface degradation by electromagnetic radiation and by atomic oxygen, perturbations in mag-

netic attitude control systems due to magnetic disturbances, and impacts by micrometeoroids and space debris (Bedingfield et al., 1996 [17]). For a more comprehensive overview of space weather effects on technological systems the reader is referred to Lanzerotti (2001) [4].

3 Space Environment and Its Variability

In general, space weather can be considered as solar-induced short-term variability of space environment. In terrestrial analogy, the quasistationary space environment corresponds to a regional climate on which the Sun generates changes representing weather phenomena. A terrestrial regional climate can be described by cyclic variation of conditions, and this is true also for the space environment. The "seasonal" changes of the near-Earth space environment due to the 11-year cycle of solar activity have been reviewed in Gorney (1990) [18]. In the following, the main characteristics of space environment and manifestations of short-term solar influence in that environment are summarised, concentrating on properties and processes significant for the effects discussed later in Sects. 4 and 5. Many of these topics are discussed in great detail elsewhere in this volume.

3.1 Space Environment

When assessing the effects of space environment on technologies relevant to the present discussion, the most important environments to be included are plasmas and energetic particle radiation and, to a lesser extent, magnetic fields, solar electromagnetic radiation, and micrometeoroids and space debris. General specifications of the space environment have been presented in ECSS (2000) [19].

Plasmas

Space plasmas are encountered in interplanetary space in the form of solar wind continuously emanating from the Sun, and in different parts of the magnetosphere and the ionosphere originating from various sources. The solar wind flows out with a speed of about 400 km/s, corresponding to a kinetic energy of just below 1 keV for protons, and has an average density of 5 cm^{-3} at 1 AU. Both the speed and the density are variable, with a 5–95 % range of cumulative probability of occurrence of 320–720 km/s and 3–20 cm^{-3}, respectively (ECSS, 2000 [19]). Two types of solar wind streams can be distinguished: the slow solar wind from the equatorial regions of the Sun, and the fast solar wind originating from coronal holes, usually located in the polar regions of the Sun. The solar wind consists mainly of protons (95 % of positively charged particles) with a small portion of doubly- ionised helium and a trace of heavier ions. Electrons are also present in sufficient numbers to make

the wind neutral. The effects of the solar wind on technologies on ground and even in near-Earth space are indirect, caused by complex interactions of the solar wind and the coupled magnetosphere-ionosphere-atmosphere system. The effects of the solar wind on the terrestrial environment have been discussed in detail in Crooker and Siscoe (1986) [20].

The ionosphere-magnetosphere system contains various plasma regimes. As in the solar wind, the plasma properties can be described by specifying particle density and particle energy, which are approximately the same for electrons and positively charged ions (protons), but change considerably with altitude and latitude under the varying strength of the gomagnetic field. At the height of few hundred kilometres (at Low Earth Orbit) the plasma is cold ($\sim$1000 K or $\sim$0.1 eV) but high density (10^3–10^5 cm^{-3}). In the plasmasphere, an extension of the ionosphere forming the inner part of the magnetosphere up to a few Earth radii, the plasma density is typically 10–1000 cm^{-3} and the mean kinetic energy of the order of 1 eV. At the plasmapause, the density drops suddenly, typically to 1 cm^{-3} at geostationary orbit (6.6 Earth radii), while the energies are high, typically in the keV range. In the plasma sheet, in the outer magnetosphere on the night side of the Earth similar conditions prevail. At high latitude polar regions, where the open geomagnetic field lines connect directly to the interplanetary magnetic field, precipitating electrons originating from the suprathermal tail of the solar wind are encountered. In this polar rain, the electron energies are in the keV range and densities a few electrons per cubic centimetre. Further on-line details of the near-Earth plasma environment can be found, e.g., in the Space Physics Textbook (http://www.oulu.fi/$\sim$ spaceweb/textbook/). As explained in Sect. 4, the magnetospheric plasma characteristics have important consequences for space systems, in particular for spacecraft surface charging.

Energetic Particles

When considering the "quiescent" space environment (i.e., the space climate), the major sources of high-energy ($\geq$100 keV) particles are galactic cosmic rays (GCR) and, within the magnetosphere, the trapped particles in the radiation belts. Galactic cosmic rays are composed of protons (83 %), ^{4}He ions (13 %), heavy ions (1 %) with significant fluxes up to the iron group of elements (Z=26– 28), and electrons (3 %). GCR are characterized by low intensities and high energies. The energy range where GCR are significant from the point of view of radiation effects on space systems extends from about 100 MeV/nucleon up to several 10 GeV/nucleon with power law spectra above a few GeV/nucleon and a spectral slope of -2.7. The integral intensity of protons above 100 MeV is $\sim$1 p cm^{-2}s^{-1}. Below about 1 GeV/nucleon, the spectra are strongly affected by solar modulation (Cane et al., 1999 [21]) leading to flattened energy spectra with a maximum at 200–300 MeV/nucleon and decreasing differential intensities at lower energies (Klecker, 1996 [22]). The strength of the modulation is solar cycle dependent with peak-level in-

tensities observable at solar minimum. The difference in proton intensities between solar cycle minimum and maximum is of the order of 10 % below 5 GeV/nucleon, but can be much larger (a factor of five) at still lower energies and for heavier ions.

The radiation belts encircle the Earth from the top of the atmosphere to the outer edges of the magnetosphere. The trapped particles are composed of energetic protons and electrons and small amounts of heavy ions (Stassinopoulos and Raymond, 1988 [23]). In orbits passing through the radiation belts, these particles pose the most significant threat to radiation-sensitive systems. Traditionally, the radiation belts are divided in two zones, the inner belt and the outer belt. The "slot" region between the inner and outer belt is a region of much lower particle fluxes. In the inner belt, protons with energies up to several hundred MeV are the most important component, but also electrons in the MeV range are present. The maximum flux of $> 10\,\mathrm{MeV}$ protons ($> 10^5\,\mathrm{p\,cm^{-2}s^{-1}}$) occurs approximately at an altitude of two Earth radii (R_E). At this altitude also heavy ions are most abundant (Mewaldt et al., 1996 [24]). Significant proton fluxes extend up to $4\,\mathrm{R}_E$ in the equatorial plane. The outer belt is dominated by electrons with energies up to at least tens of MeV. The maximum of $> 1\,\mathrm{MeV}$ electrons is at $\sim 4\,\mathrm{R}_E$ while the outer bound of the belt is at $\sim 10\,\mathrm{R}_E$. Detailed descriptions of the characteristics of the radiation belts are presented, e.g., in Vampola (1989) [25] and Daly (1994) [26].

A third source of radiation is the anomalous cosmic rays (Fichtner, 2001 [27]). The importance of this components is, however, much smaller than that of the GCR or the radiation belts, because at 1 AU the anomalous cosmic rays only have significant fluxes outside the magnetosphere and only at the time of the solar minimum. They also mainly compose of helium, nitrogen, oxygen, and some heavier ions at relatively low energies (tens of MeV/nucleon) with a limited penetrating power. Anomalous cosmic rays are the source of the heavy ions in the radiation belts (Mewaldt et al., 1996 [24]).

Finally, secondary radiation produced in interactions of very high-energy galactic cosmic ray particles in the structures of a spacecraft can cause a significant radiation background (Dyer et al., 1996 [28]). Secondary radiation consists of charged particles, neutrons, and bremsstrahlung photons. In some cases, even induced radioactivity needs to be considered.

Solar energetic particle (SEP) events are an obvious source of high-energy particles to be taken into account, but the discussion on solar particles is postponed to Sect. 3.2. Here we only note that the number of expected SEP events is dependent on the solar cycle. In terms of SEP events, the solar cycle consists of seven active years beginning two years before the year of sunspot maximum and lasting four years after the maximum (Feynman, 1990 [29]). During the rest of the solar cycle, the risk of a solar particle event is low. A general summary of solar cycle effects on space radiation environment is presented in Wilson et al. (1999) [30].

Magnetic Fields

The solar wind carries with it a magnetic field of about 5 nT that lies, due to solar rotation, near the ecliptic plane in an Archimedean spiral pattern. It has a wavy structure, which leads to a current sheet-separated sector structure with the magnetic field direction alternately towards and away from the Sun (e.g., Friedman, 1986 [31]; Russel, 2001 [32]). During nominal interplanetary conditions, the interplanetary magnetic field (IMF) has no consequences on technological systems in space, but during disturbed periods, when both the direction and strength of the IMF are highly variable, it is the most important parameter affecting the geomagnetic activity together with the velocity and density of the solar wind.

Earth's magnetic field is basically a dipole field, but is strongly distorted by the magnetised plasma of the solar wind, leading to the structure known as the magnetosphere (Otto, 2004 [33]). On the dayside, during nominal solar wind conditions, the boundary of the magnetosphere is at the distance of $10\,\mathrm{R}_E$, while on the nightside it extends to hundreds of R_E. The inclination of the dipole field is about 11° with respect to the Earth's rotation axis (Fraser-Smith, 1987 [34]) and the offset from the center of the Earth in the year 2000 epoch was 540 km. The geomagnetic field is significant for space weather effects from several respects. It controls the near-Earth plasma environment. The geomagnetic field traps the particles in the radiation belts and controls their motion. The tilt of the dipole field with respect to the Earth's rotation axis and the offset from the centre result in the South Atlantic anomaly, a region where high intensity radiation reaches exceptionally low altitudes due to the low magnetic field strength. In addition, the magnetosphere has a shielding effect on high-energy particles and determines the cut-off rigidity of particles as a function of magnetic latitude, i.e., the minimum momentum with which a particle of a certain charge can reach the atmosphere. Only through the polar cusps, where the geomagnetic field lines connect directly to the interplanetary magnetic field, can the lowest energy particles enter the atmosphere.

Solar Electromagnetic Radiation

The Sun emits electromagnetic radiation at all wavelengths from γ-rays to radio waves. The shape of the spectral distribution is such that the bulk of the solar energy lies between 150 nm and 10 µm with the maximum near 450 nm, i.e., at the visible range of wavelengths. However, the ultraviolet portion (< 300 nm) of the spectrum is the most important in determinining the effects of solar radiation on the upper atmosphere and on technological systems in space. During solar storm conditions, also X-ray fluxes are significant. The variability of electromagnetic radiation in the visible wavelength range is very small over the solar cycle. Other parts of the spectrum can be much more variable both over the 27-day solar rotation period and over the 11-year solar

cycle. For the ultraviolet part, the variability can be of the order of factor 2, and can reach orders of magnitude for flare X-rays (ECSS, 2000 [19]).

Space Debris and Micrometeoroids

Spacecraft in Earth orbit are exposed to man-made orbital debris and streams of micrometeoroids of natural origin. Space debris may become a major aspect of future space missions, but their effects are not considered in this chapter.

3.2 Solar Effects on Space Environment

The Sun emits huge amounts of mass and energy, which control the space environment. The fluctuations in the local energy production at the Sun determine the short-term conditions in the space environment. The source of these fluctuations is the changing magneticg activity of the Sun Schüssler (2004) [35]. In the following, a brief overview of the most important forms of the solar activity affecting the interplanetray and near-Earth space environment is given. A summary of the solar activity influences on space environment and effects on spacecraft is available also in Vaughan et al. (1996) [36].

Coronal Mass Ejections

Coronal mass ejections (CMEs) are enormous eruptions of magnetised plasma from the Sun (Gosling, 1997 [37]; Klimchuk, 2001 [38]). The erupting material propagates through the interplanetary medium with speeds ranging from only a few km/s to over 2000 km/s (St. Cyr et al., 2000 [39]). When the speed of a CME exceeds that of the upstream solar wind, a shock travelling ahead of the CME is produced. The CME itself contains usually coronal material or heated material from a solar filament, and carries magnetic field, which sometimes has an ordered structure of a magnetic cloud (Burlaga, 1991 [40]). The main effects of coronal mass ejections on technologies are two-fold. The direct effect is that often high-energy particles are accelerated in the coronal and interplanetary shocks associated with CMEs. Secondly, CMEs are the primary cause of large geomagnetic storms (Plunkett and Wu, 2000 [41]), which again can cause severe problems for technological systems both in space and on ground.

The sporadic occurrence of very energetic (10 MeV–10 GeV) solar particle events has been recognised as one of the most important space weather effects in the near-Earth space (Gorney, 1990 [18]). Sudden intense bursts of solar energetic particle events can reach the Earth within a few tens of minutes and highly enhanced flux levels (10^4–10^5 particles $\mathrm{cm}^{-2}\mathrm{s}^{-1}$) can last several days (Reames, 1999 [42]). These particles have ready access to the open field lines of the polar caps and can reach low Earth orbit altitudes of a few hundred kilometres. Figure 1 presents the intensity-time profiles of protons in the

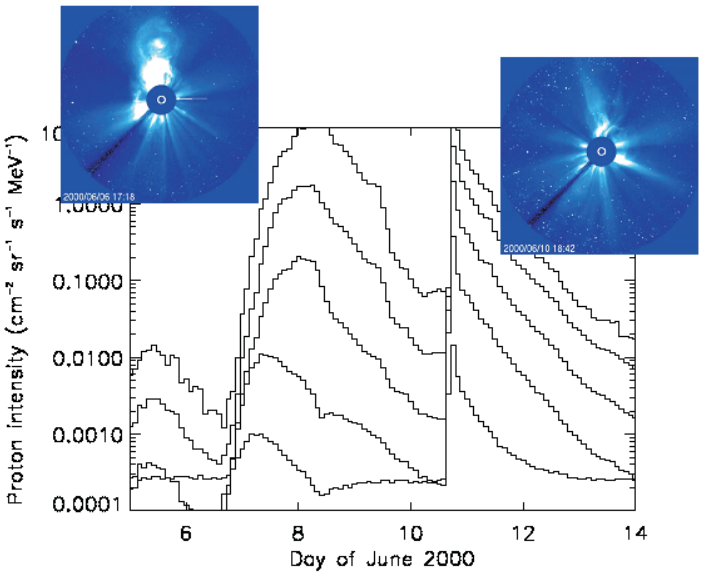


Fig. 1. Intensity-time profiles of 10–100 MeV protons in June 2000. Inset are coronagraph images of CMEs related to the particle events

range 10–100 MeV during two SEP events in June 2000. Both events were associated with a coronal mass ejection, as shown by the inset SOHO/LASCO coronagraph images. The different time profiles of the events in Fig. 1 indicate different source locations on the solar disk with respect to the observation point (Kahler, 2001 [43]). The fast rising event represents a good magnetic connection to a source near the western limb of the Sun.

Only the fastest CMEs drive shocks, and the observed particle intensities correlate with the CME speeds. If the shock is strong enough to accelerate particles still at 1 AU, a strong increase in particle fluxes can be seen when the shock reaches and passes the observer within 2-4 days after the CME launch from the Sun. Sometimes particles in this shock peak represent a major part of the total fluence of an event.

It is widely accepted that coronal mass ejections and their interplanetary counterparts are the cause of the major non-recurrent geomagnetic storms (Tsurutani and Gonzales, 1997 [44]; Webb et al., 2001 [45]). Fast CMEs ($>$ 500 km/s) can contain strong magnetic fields both in the sheath region ahead of the CME and within the CME structure itself, which interact with the geomagnetic field. The strong long-duration southward orientation of the interplanetary magnetic field ($B_z < -10$ nT, $t > 3$ hours) is a major factor in efficient coupling of solar wind energy into the geomagnetic field through magnetic reconnection in the magnetopause (Gonzales and Tsurutani, 1987 [46]). The magnetic field reconnection opens the way for the solar wind plasma

to the geospace resulting in large-scale fluctuations in plasma populations, electric currents, and magnetic fields. Some aspects of geomagnetic storms significant for the topics of this chapter will be briefly discussed later.

Solar Flares

Solar flares are an other important source of space weather effects. Here, X-rays, ultraviolet radiation, radio emission, and energetic particles play the major roles. X-rays and UV-light cause mainly indirect effects through their interaction with the Earth's atmosphere. The flare particle events are generally of lower peak intensity and shorter duration ("impulsive") than those associated with CMEs. As the interrelations of CMEs and flares are not yet fully clear (Klimchuk, 2001 [38]), so are the sites and processes of particle acceleration still discussed, particularly the origins of high-energy protons contributing to the fast-rising intensities (e.g., Cane, 1997 [47]; Klein and Trottet, 2001 [48]). It is clear, however, that flares alone can produce particle events with substantial intensities, but it is often difficult to distinguish between flare and CME shock-accelerated particles. The fast rising event on June 10, 2000 (Fig. 1), was associated with both a CME and an M5.2-class X-ray flare. In general, a good magnetic connection from the flare site to the observer is required for the flare particles to be detected. Although the effects are the same irrespective where the particles have been accelerated, it would be important for predicting the occurrence of particle events to understand their origin.

High-Speed Solar Wind Streams

High-speed solar wind streams emanating from coronal holes can drive interplanetary shocks and create intense magnetic fields when interacting with streams of lower speeds (Tsurutani and Gonzales, 1997 [44]). Similarly to CME-driven disturbances, these compressed field regions can couple to the geomagnetic field and may cause geomagnetic storms. During low solar activity, coronal holes can be relatively stable, lasting for months, and reach low solar latitudes. Therefore, at solar minimum, the high-speed solar wind streams are the dominating source of geomagnetic storms, and due to solar rotation cause recurrent storms with a 27-day pattern.

At large heliospheric distances (> 1.5 AU) the shocks associated with the high-speed streams rotating with the Sun are fully developed, and form corotating interaction regions (CIRs). CIRs are still an other source of high-energy particles (Mason and Sanderson, 1999 [49]). However, due to low intensities and steep spectra of CIR particles, the effects are relatively insignificant.

Geomagnetic Storms and Substorms

Interactions of coherent solar wind and interplanetary magnetic field structures with the magnetosphere cause major disturbances in the geomagnetic

field (Russel, 2000 [50]). The dynamic solar wind pressure can compress the magnetopause inside the geostationary orbit, leading to difficulties in spacecraft with magnetically controlled guidance systems. The most important consequences for space systems, however, result from the enhanced plasma and particle environment.

Substorms and geomagnetic storms are manifestations of magnetospheric response to the rate of magnetic reconnection at the magnetopause. Occurrence of substorms does not require particularly disturbed interplanetary conditions, and on the average there are one to several isolated substorms during a day. In the substorm growth phase, energy dissipated by the solar wind is stored in the magnetotail, and is explosively released in the expansion phase, leading to acceleration of electrons in the keV-range (Baker, 1996 [51]). As a consequence, dense clouds (1–10 cm^{-3}) of energetic electrons are injected from the tail region towards the inner magnetosphere.

Geomagnetic storms are initiated, when enhanced energy transfer from the solar wind into the magnetosphere, in response to CMEs or high-speed solar wind streams, leads into intensification of the ring current. Geomagnetic activity controls the trapped energetic particle environment, in particular the outer radiation belt electron fluxes. During geomagnetic storms, relativistic electron fluxes are strongly and rapidly enhanced in the outer belt (Baker et al., 1997 [52];Baker et al., 1998 [53]). Extended geomagnetic activity can cause 2-3 orders of magnitude increases in the > 100 keV electron peak fluxes lasting for several days.

An extreme effect of geomagnetic storms on the Earth's radiation environment is the creation of new radiation belts, i.e., filling with high fluxes of particles for a duration of many weeks to months a region in space, where no significant particle populations existed before. This is a rare incident, but was clearly observed in March 1991 (Mullen et al., 1991 [54]). CRRES spacecraft detected the formation of a second peak in the inner proton belt immediately following the sudden storm commencement on 24 March 1991. The birth of the new radiation belt was attributed in Mullen et al. (1991) [54] to the injection of high-energy protons by the solar-initiated shock accompanying the storm sudden commencement deep into the magnetosphere ($2.5\,\mathrm{R}_E$). Subsequently, weaker proton belt formations related to other solar proton events were reported (Gussenhoven et al., 1994 [55]), as well as long-duration enhancements of high- energy (> 2 MeV) electrons in the radiation belt slot region associated with geomagnetic storms (Gussenhoven et al., 1996 [56]; Baker, 2000 [57]).

3.3 Space Environment Models for Effects Calculation

Already in the design phase of a spacecraft or a space instrument it is necessary to evaluate the environment encountered during the mission and the effects this environment might have on the system. The basic models that can be used for these purposes are the NASA radiation belt models AE-8 for

electrons (Vette, 1991 [58]) and AP-8 for protons (Sawyer and Vette, 1976 [59]), the JPL-91 solar proton fluence model (Feynman et al., 1993 [60]; Feynman et al., 2002 [61]), and CREME96 (Tylka et al., 1997 [62]) for modelling galactic and anomalous cosmic rays. The main deficiency of the radiation belt models AE-8 and AP-8 is that they only describe the quasistationary radiation belt environment without any dynamic features other than providing separate models for the solar minimum and maximum. The general problem in the most extensively used near-Earth environment models is that space weather effects on the environment are not taken into account. Some improvements have been gained by applying the CRRES results in the modelling (Gussenhoven et al., 1996 [63]). Recently, a new model has also been developed for predicting cumulative solar proton fluences and worst case solar proton events as a function of mission duration (Xapsos et al., 1999 [64]). The most common environmental models have been reviewed in detail in Barth (1997) [65].The problems encountered in predicting various space weather effects have been considered in Feynman and Gabriel (2000) [66]. For using many of the models listed above, as well as for models for calculating the actual effects caused by the environments, the reader is directed to the SPENVIS online system providing tools for analysis of the space environments and their effects (http://www.spenvis.oma.be/spenvis/).

4 Plasma Effects

Technological systems in space will interact with the plasma environment surrounding them. Accumulation of charged plasma particles on spacecraft surfaces leads to high potentials of those surfaces relative to their surroundings. Subsequent electrostatic discharges are severe threats to spacecraft. Charge buildup affecting operational characteristics of various instruments is another concern. A third type of a plasma effect is degradation of thermal and optical surface properties in long-term exposures to space plasmas.

4.1 Surface Charging

In a comprehensive study of space environment effects on space systems (Koons et al., 1999 [67]), surface charging was found to be the most significant cause of spacecraft anomalies leading to mission failures. Surface charging is the process of electric charge accumulation on surfaces exposed to space environment. It is produced by interactions between satellite surfaces and space plasma, geomagnetic fields, and solar electromagnetic radiation (Leach and Alexander, 1995 [68]). Due to changing environment, the accumulated charge is never steady.

Two types of surface charging can be distinguished (Purvis et al., 1984 [69]). Absolute charging occurs, when the entire spacecraft potential relative to the ambient plasma is changed uniformly. Differential charging means

that parts of the spacecraft are charged to different potentials relative to each other. In this case, strong local electric fields may exist. Absolute charging, which in itself is not generally detrimental, occurs typically in eclipse (in the shadow of the Earth) very rapidly, in fractions of a second. Differential charging is a relatively slow process (minutes), and strongly depends on the dielectric material properties. Non-uniform material properties can lead to high potential differences between various parts of a spacecraft. Environmental factors, such as exposure to sunlight and anisotropic plasma fluxes, greatly affect the level of differential charging.

The potential of a body in space is determined by the balance between various charging currents. At equilibrium, all currents must sum to zero. The potential at which equilibrium is achieved is the potential difference between the spacecraft and the space plasma ground. A simple approximation expressing this current balance can be written as (Purvis et al., 1984 [69])

$$I_e(V_s) - [I_i(V_s) + I_{se}(V_s) + I_{si}(V_s) + I_{be}(V_s) + I_{ph}(V_s) + I_a(V_s) + I_s(V_s)] = I_t(V_s)\ . \tag{1}$$

When equilibrium has been reached, the satellite surface potential V_s has a value giving the total current $I_t = 0$. In (1), I_e and I_i are the currents due to incident plasma electrons and ions, respectively. Since the density and kinetic energy of electrons and positively charged ions are approximately the same, and because the mass of the electrons is much smaller than that of the ions, the electrons move much faster, and the negative electron current is greater than the positive ion current. The currents I_{se} and I_{si} are due to secondary electrons emitted from the spacecraft surface by the collisions of the plasma electrons and ions. At low energies (in the eV range), the secondary emission ratio by the dominating electron component exceeds unity. At electron energies in the keV range, the ratio drops below unity, and in such an environment the surface begins to charge. For shadowed surfaces, the equilibrium potential is often determined by a balance between incident primary electrons and the resulting secondary emission. The current I_{be} in (1) represents backscattered plasma electrons, which leave the material with somewhat lower energies than they had upon entering. A spacecraft surface charging negatively by I_e, prevents low-energy electrons from reaching it. A sheath region is created around the surface repelling electrons and attracting the positive ions. Equilibrium is reached when the sheath region has grown to a sufficient extent to balance the currents due to negative and positive plasma species (Vampola, 1989 [25]). A very important contribution to the equilibrium comes from the photoelectric current, I_{ph}, due to the solar UV radiation. In sunlight, the photoelectric current from a surface is usually much greater than the plasma current. Photoemission from the extreme ultraviolet wavelength range ($< 200\,\mathrm{nm}$) is the most important, since in that region the solar spectrum still has significant energy and photoelectric yields of many materials are large (Whipple, 1981 [70]). Therefore, in sunlight the equilib-

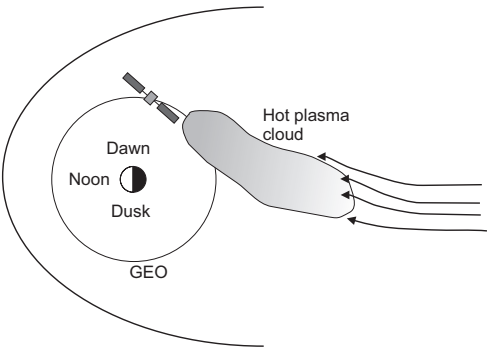


Fig. 2. Schematic view of a high-temperature plasma cloud drifting towards GEO

rium is controlled by emission and reattraction of photoelectrons. The last two terms in the left-hand side of (1) are the current from artificial (active) sources (I_a) and the current along the differentially charged spacecraft surfaces and through the structures (I_s). Based on (1), it can be shown (Purvis et al., 1984 [69]) that while the spacecraft is in eclipse the equilibrium potential (in Volts) will be $V_s \approx -T_e$, where T_e is the environment electron temperature in electronvolts. This relation is valid for environments, where electrons have sufficient energies ($> 1\,\mathrm{keV}$) such that secondary electron production is not anymore significant, and shows that in eclipse the spacecraft potential is approximately numerically equal to the plasma temperature.

In sunlight, the photoelectric current usually balances the absolute spacecraft potential to a few volts positive (Grard et al., 1983 [71]). In regions where the cold plasma density is low, the possibility exists that surfaces can charge to very high potentials. Magnetospheric activity plays a decisive role in spacecraft charging. During substorm activity, high-temperature (up to tens of keV) plasma clouds are generated and transported from the geomagnetic tail towards the Earth (Fig. 2). These clouds can encounter satellites at geostationary orbit (GEO). In eclipse, in the absence of photoelectric current due to sunlight, surfaces may charge to the levels of tens of kV. The dawn-side drift of the negatively charged electrons in the geomagnetic field when moving towards the Earth, leads to the classic pattern of most spacecraft anomalies related to surface charging occurring in the midnight-to-dawn local time sector (Gubby and Evans, 2002 [16]; Lam and Hruska, 1991 [72]). This pattern appears to be an immediate response to the magnetic perturbations (Lam and Hruska, 1991 [72]). When exiting eclipse, the sunlit surfaces charge positively due to photoelectron emission, while some parts of the spacecraft structure may remain shadowed and stay in high negative potential. As well, different surface materials discharge at different rates creating large differential potentials.

In low-Earth orbit, satellites experience high-density low-energy plasma, and surface charging is not usually a concern. However, in cases when the

satellite velocity is greater than the ion velocity, but slower than the electron velocity, a wake effect results. In such a situations, positively charged ions can only impact ram surfaces, while fast electrons are capable of impinging on all surfaces. The wake effect may lead to differential charging of surfaces.

The effects of surface charging on spacecraft systems come from discharge arcing. Whenever the charge buildup generates an electric field exceeding a breakdown threshold, charge will be released. The discharge may occur through dielectric breakdown (punch-through) or between surfaces (flash-over). As a result, currents are flowing in spacecraft structures and broadband electromagnetic fields are produced coupling into the electronics. The effects include physical spacecraft surface damage and degradation, and operational anomalies, such as telemetry glitches, logic upsets, component failures, and spurious commands (Leach and Alexander, 1995 [68]). A well-documented example of surface charging effects is the switching anomalies of the maritime European communications satellite Marecs-A (Capart and Dumesnil, 1983 [73]). A list of spacecraft charging anomaly events is given in Leach and Alexander (1995) [68], but not distinguishing between surface and internal (see Sect. 5.1) charging.

4.2 Plasma Effects on Instruments

Due to the sheath region surrounding a charged spacecraft in equilibrium, the lowest-energy particles may be repelled and cannot reach the spacecraft. This may become a problem for an instrument trying to measure the low-energy plasma properties. It was this interfering effect of spacecraft charge on low-energy particle and electric field experiments that stimulated the investigation of methods of spacecraft potential control (Whipple, 1981 [70]). Instrument bias with respect to plasma ground may distort the energy distribution of incidedt ions. Perturbations of ion trajectories also lead to changes in angular resolution and sensitivity. In general, the response of a plasma instrument will change compared to the pre-flight calibration.

Sputtering of surfaces, i.e., physical removal of surface atoms, due to considerable ion kinetic energy may degrade the quality of optical and X-ray mirrors. Sputtering products may be ionised by solar radiation while still near the spacecraft and attracted to negatively charged surfaces contributing to contamination.

Solar ultraviolet radiation and plasma interactions can cause dust generation and shedding of surfaces. Particulate contamination on a payload can become detached when it is exposed to an ionospheric plasma (Goree and Chiu, 1993 [74]). Once the dust is released, it can remain in the immediate environment of the spacecraft and interfere with the operation of optical instruments by scattering light.

A specific type of instrument interference is the subauroral red arc (Vampola, 1989 [25]). Low energy ($> 20\,\mathrm{keV}$) protons from the ring current produced by magnetic storms can precipitate into the atmosphere producing a

diffuse red area at intermediate magnetic latitudes, which potentially interfere with optical instruments on low-altitude satellites.

5 Radiation Effects

High-fluxes of high-energy electrons produce internal charging of spacecraft components, which has led to several operational anomalies. Probably the most frequently reported radiation effects on space systems are, however, various types of Single Event Effects (SEE) caused by single particle strikes in spacecraft electronics. Total ionising damage due to high radiation doses and displacement damage in the bulk of semiconductor devices are also of major concerns in space. In addition to electronic components, these damage mechanisms have significant effects on solar arrays and semiconductor detectors. Finally, high-energy radiation creates sensor background and interference in scientific instruments.

5.1 Internal Charging

Internal charging refers to the process where impinging charged particles have sufficient energies to penetrate slightly below the surface of a dielectric material and are trapped. Figure 3 shows the range of electrons and protons in aluminium. It is noticed that electrons with energies above about 100 keV can penetrate spacecraft surface metallisation and other thin layers (of the order of 100 µm) of materials exposed to space. These penetrating electrons can become embedded inside insulating materials such as thermal blankets, coaxial cables or circuit boards. Electrons with energies of 1 MeV can already pass through a 2 mm-thick layer of aluminium and correspondingly thicker

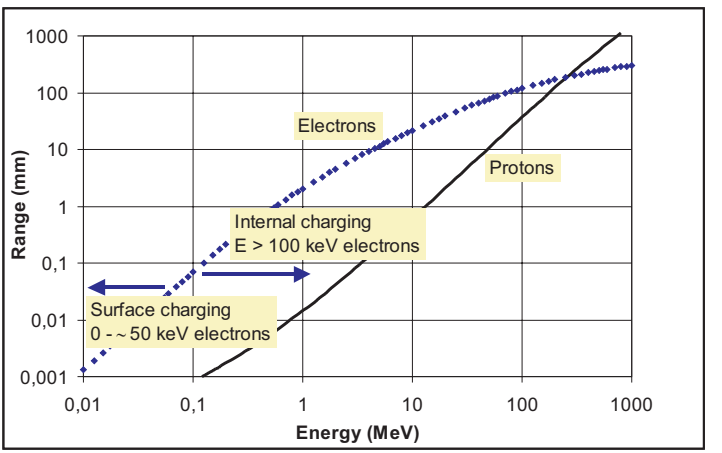


Fig. 3. Range of electrons and protons in aluminium

layers of less dense materials. Protons with similar penetration power require much higher energies, and because particle intensities generally are steeply falling with energy, protons are not usually considered as a cause of internal charging problems.

The level of internal charging depends on the environment, the shielding thickness of the spacecraft, and the characteristics and shape of the charged material (Leach and Alexander, 1995 [68]). When the rate of energetic electron deposition inside an insulator or an isolated conductor is greater than the rate at which the charge leaks out, the charge is building up and the local electric field is increasing in magnitude. In the presence of a sustained flux of high-energy electrons, the electric field can exceed the breakdown threshold of the dielectric and an arc discharge occurs. The absence of neutralising effect from space plasma ions or photoemission allows internal charging to proceed slowly with the charge buildup taking for hours or days.

In its simplest form, a charging dielectric can be examined in a one-dimensional, planar approximation at a fixed depth. The time dependence of the electric field E is then described by the differential equation

$$\varepsilon(dE/dt) + \sigma E = J \ , \tag{2}$$

where ε is the dielectric constant, and σ and J are the conductivity of the material and the current density, respectively, assumed to be independent on time. The solution of (2) is

$$E(t) = E_o exp(-\sigma t/\varepsilon) + (J/\sigma)[1 - exp(-\sigma t/\varepsilon)] \ , \tag{3}$$

where E_o the initial value of the electric field. Equation (3) shows that the electric field increases with a time constant $\tau = \varepsilon/\sigma$, which typically has values in the range 10–10^4 s. After a long time, the elctric field reaches the value J/σ. If this exceeds the dielectric strength, a breakdown occurs.

The basic environmental cause of internal charging are electrons accelerated in the magnetosphere during extended intervals of geomagnetic activity. As discussed in Sect. 3.2, electron fluxes in the outer radiation belt are highly variable, with increases caused by major geomagnetic storms as great as four orders of magnitude in short time scales. Typical decay constants of outer zone electrons are of the order of 10 days. As a consequence of magnetic storms changing the large scale morphology of the geomagnetic field, particles experience inward radial diffusion from the geomagnetic tail and are accelerated to higher energies. In addition to radial diffusion, geomagnetic storms also cause pitch angle scattering of particles. Therefore, previously stably trapped particles can be perturbed to reach lower altitudes, and spacecraft which normally are below the trapped radiation zones may suddenly find themselves in large fluxes of energetic electrons at midlatitudes (Vampola, 1989 [25]). The effect of magnetic storms at geosynchronous orbit is to temporarily depopulate energetic electrons. Major storms, however, produce high fluxes of energetic electrons deeper in the magnetosphere, which rapidly dif-

fuse back out producing a harder spectrum with enhanced levels of > 1 MeV electrons at GEO-region (Vampola, 1989 [25]; Baker et al., 1998 [53]).

In assessing internal charging, the time-integrated flux of particles (i.e., the fluence) is the essential quantity. Various limits for electron fluences for discharges to occur have been reported (e.g., Gorney, 1990 [18]; Wrenn, 1995 [75]). A rule of thumb used in issuing warnings to satellite operators for potentially damaging conditions is that the daily flux of > 2 MeV electrons should exceed 3×10^8 $\mathrm{cm}^{-2}\mathrm{sr}^{-1}$ for 3 consequtive days or that the flux is greater than $10^9\,\mathrm{cm}^{-2}\mathrm{sr}^{-1}$ for a single day. The energy distribution of the incident electrons is also a critical factor (Wrenn, 1995 [75]).

As discussed in Sect. 4.1, surface charging events tend to occur in the midnight-to-dawn local time sector. This is not the case in general for spacecraft anomalies attributed to internal charging. Anomalies recorded on the DRA δ satellite were analysed in Wrenn (1995) [75] and it was shown that there was no local time dependence. Meteosat-3 anomalies were studied in Rodgers et al. (1998) [76], and the conclusion was that events occurring at times of highest fluxes of 43–300 keV electrons occurred almost exclusively between 3 and 9 hours local time. Those occurring at times of lower fluxes occurred at any local time. Both groups of anomalies were, however, attributed to internal charging. Internal charging tends to occur at all local times, partly because penetrating electrons are distributed in the magnetosphere more uniformly than hot electron plasma clouds responsible for surface charging, and partly because internal charging is a cumulative phenomenon. The classic midnight-to-dawn pattern can exist, however, during the most active periods.

As in the case of surface charging, also internal charging results in arc-discharging and the effects are similar (Leach and Alexander, 1995 [68]). Internal charging effects can, however, be more severe because of the usually closer proximity of the discharge area to the sensitive elements (Fig. 4). Weak discharges can cause spurious signals, but severe discharges can physically damage semiconductors. Both coaxial cables and printed circuit boards have been shown to be susceptible to internal charging damage when irradiated with electron beams (Wrenn, 1995 [75]). Internal charging was studied in detail together with simultaneous comprehensive measurements of environment conditions in the CRRES mission (Frederickson, 1992 [77]). Based on the CRRES results it was conclude in Gussenhoven et al. (1996) [56] that most of the environmentally induced spacecraft anomalies result from internal charging and not from surface charging or other radiation effects.

5.2 Total Ionising Dose

Energetic particles and photons passing through matter lose energy through ionisation of the medium, i.e., producing electron-hole pairs by stripping off electrons from the atoms of the medium through electromagnetic interaction. In this process, the valence band electrons in the material are excited to the

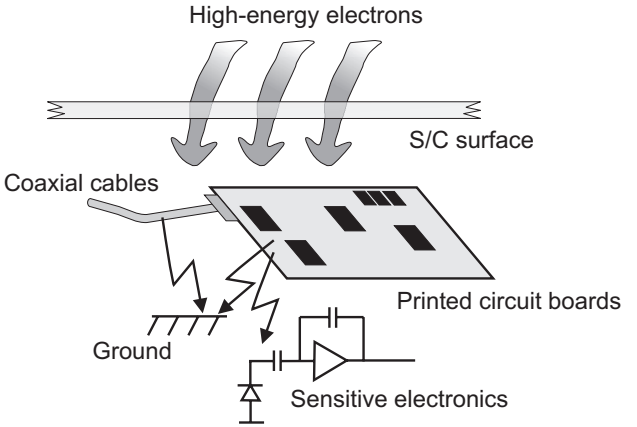


Fig. 4. Illustrative examples of discharges following internal charging of coaxial cables and printed circuit boards

conduction band, where they are highly mobile, if an electric field is applied. The deposition of energy by a charged particle in a material by means of ionisation is conventionally termed "dose", and is measured as the energy deposited per unit mass. The SI unit of dose is Gray although the old unit, rad, is till commonly used (1 Gy = 100 rad).

The ionisation energy loss rate is given by the Bethe-Bloch formula as

$$-(dE/dx) \propto (nz^2Z)/(mv^2) \; . \tag{4}$$

Here z, m, and v are the charge number, mass, and velocity of the incident particle, respectively, and n and Z are the number density and charge number of the medium. The Total Ionising Dose (TID) is obtained as the integral of the particle flux and the ionisation energy loss rate. The main contributions to TID in space come from trapped protons and electrons, and from solar protons during CME or flare events. In particular, protons with relatively low energies, which still are capable of penetrating spacecraft surfaces (c.f., Fig. 3), are important due to their high fluxes and due to the inverse-square dependence of the ionisation loss on velocity. In some circumstances, doses from bremsstrahlung photons produced by high-energy electrons stopping in spacecraft structures can be significant. In spite of the z^2-dependence of the energy loss, doses from trapped heavy ions, on the other hand, are insignificant, because they are easily absorbed on spacecraft surfaces due to their moderate energies. As well, contributions from galactic cosmic ray protons and heavy ions are small because of their low intensities.

Tools have been developed for calculating the total ionising dose for a specific radiation environment and amount of shielding of a sensitive part. In principle, the following basic steps have to be included. By specifying the mission, an orbit for the spacecraft can be selected. When this has been done,

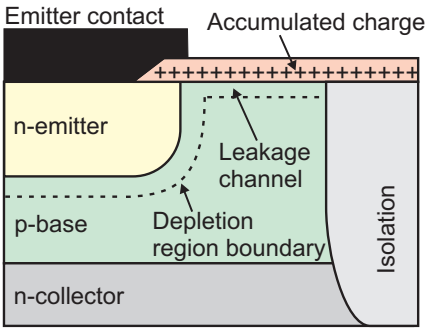


Fig. 5. Generation of a leakage channel in an NPN transistor by accumulated positive charge after a high total radiation dose

the environmental models briefly described in Sect. 3.3 can be used for calculating the fluence (mission-integrated flux) as a function of energy for protons and electrons. TID can then be calculated for simple device geometries by folding pre-computed datasets of radiation transport results with this spectrum (ECSS, 2000 [19]). As the result, dose at a given depth is obtained. Examples of radiation transport results and dose-depth curves for various missions are given in ECSS (2000) [19]. For more detailed assessment of radiation effects, precise simulations are needed. Recently, a multilayer shielding simulation software tool has been developed for this purpose (Lei et al., 2002 [78]). The Spenvis system (Sect. 3.3), is an online tool, which allows all the steps briefly described above to be implemented and total ionising doses easily calculated.

The number of electron-hole pairs produced in a medium by a particle with a certain ionisation energy loss depends on the energy required for creating a pair. In silicon, this energy is 3.6 eV and in silicon dioxide (SiO_2), commonly used as an insulator in silicon devices, it is 18 eV. Therefore, ionisation produces electron-hole pairs in SiO_2 at the rate of 8×10^{14} pairs/Gy cm^3. The electrons produced have high mobility, and are quickly swept away by internal electric fields in biased semiconductor devices. The holes have much lower mobility. Some fraction of the holes will be trapped in the bulk silicon dioxide and form recombination centres. An other part of the holes is transported to the Si/SiO_2 interface, where they will be trapped and act as fixed positive charge, and disturb the labile atomic bonds at the insulator-semiconductor interface forming new interface states (Holmes-Siedle and Adams, 2002 [79]). The trapped positive charge and new interface states are the major TID effects in both bipolar and metal-oxide-semiconductor (MOS) devices. As an illustration, Fig. 5 shows a situation where positive trapped charge in an NPN transistor creates a leakage channel increasing the base current component, which is injected from the emitter to the base region, but does not reach the collector. As a result, the gain of the transistor is degraded.

In general, the effect of TID is a change in the static and dynamic response of an electronic component. In MOS devices, important TID effects are threshold voltage shifts (e.g., Srour and McGarrity, 1988 [80]), which lead to increased stand-by and operating currents, degradation in input logic levels, reduction in noise margin, and increased propagation delay. Ultimately, the total ionising dose can cause a complete functional failure of a component.

In practice, the total ionising dose effects are much more complicated than the elementary processes of hole trapping in the SiO_2 bulk and Si/SiO_2 interface described above. Oxide traps gradually anneal with time. Interface traps are unstable and can be positively or negatively charged depending on the Fermi level position with respect to the state energy location (Edmonds et al., 2000 [81]). The density of bulk charge and interface states depend on device history, such as manufacturing process, dose and dose rate, applied voltage and temperature (Holmes-Siedle and Adams, 2002 [79]).

In devices employing thick oxide layers, an enhanced low-dose-rate sensitivity (ELDRS) may be important. ELDRS essentially means that the amount of bulk oxide charge in a device irradiated at a low dose rate (e.g., 0.01 Gy(Si)/s) can become higher than in a device receiving the same dose at a higher rate (Pease, 1996 [82]). Such a behavior has important implications for space applications, which commonly encounter much lower dose rates than components under test conditions. ELDRS appears to be connected with the dynamics of charge motion in thick oxides (Edmonds et al., 2000 [81]), but the mechanism is complex and device response varies greatly depending on the manufacturer and the technology.

Total ionising dose is a long-term, cumulative failure mechanism. It can be described in terms of a mean time-to-failure, which is the amount of mission time until the component has encountered enough dose to cause a failure. The total charge created by a single particle in a device is too low to cause any significant damage. Consequently, total ionising dose degradation results only from a large number of interactions of charged particles in a sensitive volume of a device. Exceptions are, however, miniaturised devices, which can be susceptible to microdosimetry effects (Oldham et al., 1993 [83]). Microdosimetry effect is an occurrence, where the total dose deposited by a single energetic heavy ion can cause permanent effects.

5.3 Displacement Damage

Displacement damage is the main degradation mechanism in space for certain semiconductor devices, such as solar cells, charge-coupled devices, and other photonic devices. Displacement damage is commonly described in terms of Non-Ionising Energy Loss (NIEL), which refers to that part of absorbed energy not going into the ionisation of the medium. NIEL represents the kinetic energy transferred in the interactions of incident particles to the entire atoms of the medium, and typically is orders of magnitude smaller than the ionisation energy loss in a material. Typical non-ionising processes are elastic

and inelastic scattering displacing atoms in the crystal lattice. The degree of displacement damage is proportional to the lattice deffect density. As the total ionising dose, displacement damage is a long-term, cumulative effect with lattice defect density increasing with irradiation.

The particles contributing in displacement damage in space include protons, electrons, and secondary neutrons produced in nuclear interactions of high-energy protons in spacecraft structures. Displacement of an atom from the silicon lattice requires 21 eV of energy. Because an electron is very light compared to a silicon atom, the threshold of displacement for electrons is high, about 220 keV (Bräunig and Wulf, 1994 [84]). The most significant contribution to displacement damage comes from protons in the energy range 10–200 MeV, which is sufficient for penetrating into the sensitive parts through the surface of a spacecraft (Hopkinson, 1984 [85]). Outside the inner radiation belt, the high-intensity bursts of solar protons are the most serious concern.

Depending on the species and energy of incident particles, the defects in silicon lattice occur either as isolated point defects or as defect clusters, which are highly disordered regions in the lattice. When a displaced atom moves into a non-lattice position, it leaves behind a vacancy and itself creates a defect referred to as an interstitial. Together the vacancy and interstitial form a Frenkel pair. Two adjacent vacancies form a divacancy. Together with impurity atoms, vacancies and insterstitials can form defect-impurity complexes. Most of the created vacancies and interstitials recombine, but those vacancies that remain are mobile and often combine with impurity atoms or cluster with other vacancies forming long-lived, immobile complexes (Holmes-Siedle and Adams, 2002 [79]).

The complexes are usually electronically active, having energy levels in the bandgap between the valence and conduction bands of the semiconductor. It is these energy states that have a major impact on the electrical behaviour of semiconductor devices. The resulting effects include thermal generation of electron-hole pairs (carrier generation), removal of electron-hole pairs (recombination) by successive capture of charges of opposite signs in the defect centre, temporary trapping of charge carriers, reduction of donor or acceptor equilibrium concentration by majority carrier removal, and tunneling of carriers through the potential barrier between the valence and conduction bands (Srour and McGarrity, 1988 [80]). The consequences seen in device parameters depend on the device type. Carrier generation through radiation-induced defects leads to increased leakage currents (thermal dark currents) in silicon devices. Recombination reduces minority carrier lifetime and is the mechanism for reduction of short circuit current in solar cells (Holmes-Siedle and Adams, 2002 [79]). Trapping of majority carriers leads to carrier removal, and trapping of minority carriers to a loss in charge transfer efficiency in charge-coupled devices (Hopkinson, 1984 [85]).

As an example of displacement damage in two types of silicon devices, Fig. 6 shows the long-term behaviour of the SOHO solar array relative output power and the leakage current of one of the silicon particle detectors of the

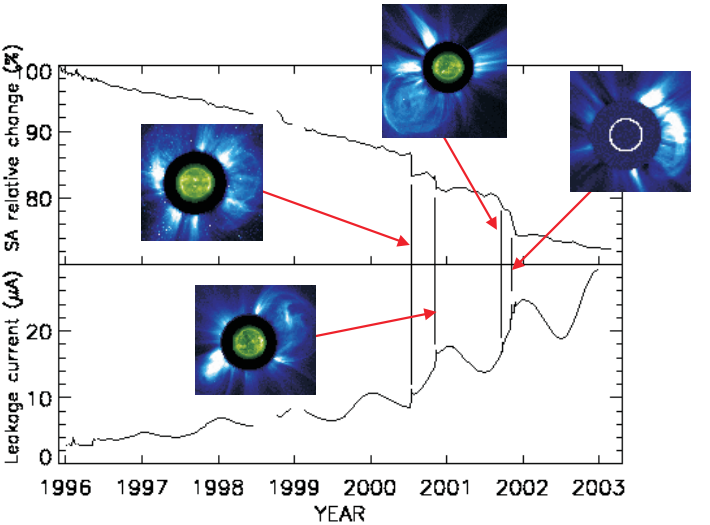


Fig. 6. An example of degradation of the solar array output power (upper panel) and silicon detector leakage current (lower panel) due to displacement damage

SOHO/ERNE experiment. Four major space weather events are indicated, with inset images from the SOHO/LASCO and SOHO/EIT. Degradations in solar array power output and increases in detector leakage current are evident coinciding with the solar events, and are most probably due to displacement damage caused by high-energy protons in the devices. The wavy structure of the leakage current (Fig. 6, lower panel) is due to a temperature effect, and the radiation damage is seen as spikes in the current.

In studies of displacement damage, the concept of the non-ionising energy loss is important, because it appears to be proportional to the total number of defects created. To a first approximation, this number is independent on the interaction details, and thus also independent of the type of particles with equal non-ionising energy losses in a device. Device parameters also degrade by the same amount for any particle energy depending only on the absorbed damage dose deposited, which is obtained as a product of particle fluence and NIEL at each energy (Summers et al., 1995 [86]). For calculation of the NIEL, the information required is the differenetial cross section $\mathrm{D}\sigma/\mathrm{D}\,\Omega$ for atomic displacement, the average recoil energy T of the target atom, and the form of the Linhard partition factor L giving the fraction of transferred energy that is non-ionising. NIEL can then be written as an integral over the solid angle as (Summers et al., 1995 [86])

$$NIEL(E) = (N/A) \int_{\theta_{min}}^{\pi} \frac{d\sigma(\theta, E)}{d\Omega}\, T(\theta, E)\, L[T(\theta, E)]\, d\Omega\ , \qquad (5)$$

where N is Avogadro's number, A is the mass numebr of the medium, and θ_{min} is the scattering angle (in the centre of mass system) for which the recoil energy equals to the threshold for atomic displacement. As an example, the NIEL curve for protons can be found in ECSS (2000) [19].

The assumption that displacement damage only depends on the non-ionising energy loss greatly reduces the amount of testing required, because the damage can be scaled according to the NIEL. In the ideal case, one experimental measurement is sufficient, e.g., by using 10 MeV protons. When the damage factor K_D relating the amount of degradation of a certain parameter to the absorbed damage dose at the test energy is known, the damage in a specified mission can be calculated as (Honpkinson et al., 1996 [87])

$$\text{Mission Damage} = \int_0^\infty K_D \, \frac{NIEL(E)}{NIEL(10\,MeV)} \, \frac{d\Phi}{dE} \, dE \, , \tag{6}$$

where $d\Phi/dE$ is the differential fluence spectrum. The mission damage is therefore obtained by multiplying the equivalent 10 MeV proton damage fluence by the damage factor. The assumption of displacement damage proportionality to NIEL is not, however, strictly valid for all particles and all types of devices. Complications in this formalism have been discussed in, e.g., Edmonds et al. (2000) [81] and Summers et al. (1995) [86].

5.4 Single Event Effects

In general terms, a Single Event Effect (SEE) is defined as a radiation induced observable effect in microelectronics circuits caused by a single charged particle losing energy by ionisation in a small sensitive target. Single event effects are instantaneous, and are evaluated as a probability of occurrence within a known mission time, which depends on the energetic particle environment and device characteristics. There are many different types of SEEs with the two main categories of soft errors, which are non-destructive and device operation is recoverable, and hard errors, which are potentially destructive and cause permanent functional effects. As yet, however, there is no generally accepted terminology, which could be used for categorising all single event effects. The most frequently discussed single event effects are Single Event Upsets (SEUs) and Single Event Latchup (SEL). SEUs appear as bitflips in digital circuits and as transient signals in analogue circuits. SEL is a potentially destructive condition involving parasitic elements in a semiconductor component. SEU, SEL, and several other types of SEE are described in more detail in, e.g., Edmonds et al. (2000) [81]

The important parameter controlling the occurrence of SEE is the ionisation energy loss rate of incident particles. The rate of energy loss in ionisation is given by (4), and in SEE studies it is commonly referred to as the Linear Energy Transfer (LET), the energy deposited in ionisation per unit path length. When passing through a device, a charged particle leaves behind a

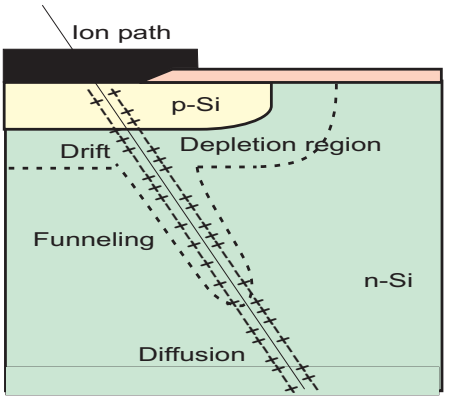


Fig. 7. Ionisation path of a charged particle in a silicon p-n junction (following Holmes-Siedle and Adams (2002) [79])

column of mobile charge carriers creating a conductive path through the circuit. Figure 7 illustrates the situation immediately after a charged particle has passed through a reverse-biased p-n junction of a silicon device. Under the influence of the electric field across the depletion region, the electron-hole pairs are quickly separated and collected on the opposite sides of the junction. This results in a short current pulse. The passage of the ionising particle also modifies the original electric field in the depletion region extending it along the path of the particle (funneling, see Fig. 7). The charge collected rapidly by this temporary electric field may significantly contribute to the prompt current pulse. If the current pulse is large enough and lasts for a long enough period, it can lead to a change of state in a digital circuit or erratic behaviour in an analogue circuit. In some circumstances a permanent current path may be created, leading to a single event latchup. An additional, slow current component comes from the charge collected by diffusion before recombination, and can be important in devices with slow response times.

Recalling the z^2-dependence of the energy loss rate in (4), it is clear that the penetrating, high-energy heavy ions of galactic cosmic rays are an important source of single event effects. However, solar energetic particle events pose the most extreme environment for SEE, particularly for spacecraft in interplanetary space and in polar orbits, where energetic particles have an easy access in the absence of a strong magnetic shielding. Because solar flare ions are of much lower energies than galactic cosmic rays, but often have relatively high intensities, it is important to take into account the exact shielding distribution of electronic components in estimating the single event effects (Kuznetzov and Nymmik, 1996 [88]). Protons typically are not ionising enough to directly cause a SEE. Instead, the process in this case goes through a nuclear spallation reaction, where a proton interacts with a nucleus in the sensitive region of a semiconductor device, and the heavy reac-

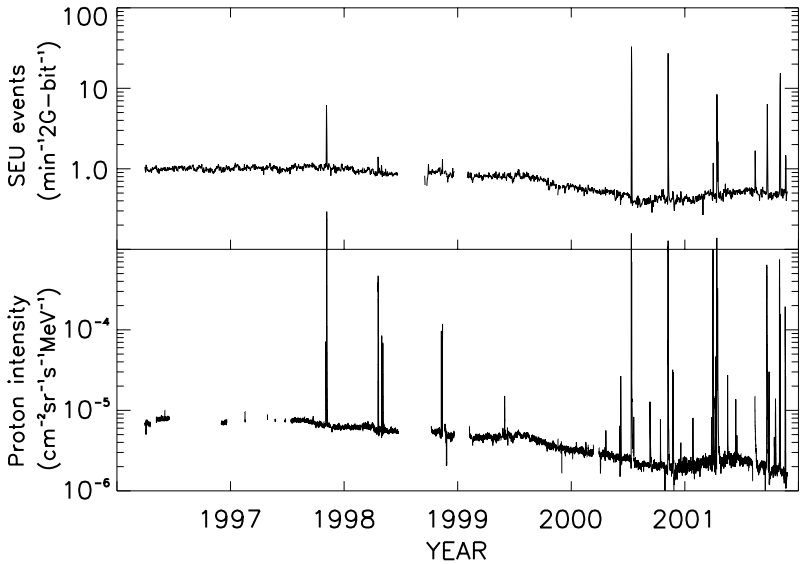


Fig. 8. Observed SEU rate in the SOHO solid state recorder (upper panel) and measured intensity of 110–140 MeV protons (lower panel)

tion products create the electron-hole pairs. With the increasing sensitivity of microelectronics to SEE, protons have become an important environmental factor. Low-Earth orbit spacecraft typically experience the highest rates of single event upsets in the neighborhood of the South Atlantic anomaly, where high fluxes of energetic protons reach low altitudes. In (Gussenhoven et al., 1996 [56]) it was concluded that most of the SEUs seen in CRRES came from high-energy protons via nuclear interactions. An example of correlation of the rate of single event upsets and the high-energy solar proton flux is presented in Fig. 8. The figure shows the observed SEU rate in the SOHO solid state recorder from the beginning of April 1996 till the end of November 2001 together with the SOHO/ERNE 110–140 MeV proton flux. The highest SEU rates clearly correlate with solar proton events. In the background rate, the solar modulation of galactic cosmic rays is also visible.

The SEU behaviour of a device is characterised by a critical charge and cross section. The critical charge is the minimum amount of charge required to cause a SEU. When the energy required for creating an electron-hole pair is known (3.6 eV for Si), the critical charge can be converted to a threshold energy, which can be used to calculate the threshold LET for a certain path length in the sensitive region. The cross section for causing an upset is determined by the geometry of the sensitive region, which often is the depletion region of a device. The upset rate can be calculated by integrating the cross section, the path length distribution through the charge-collecting region, and the directional energy distribution of ions (Robinson et al., 1994

[89]). The calculation can be greatly simplified by combining the information contained in the energy spectra of individual ions in one function, the LET spectrum (Heinrich, 1994 [90]). Based on the known relation between the LET and energy of each ion, the differential energy spectra of all particles can be converted into one differential LET spectrum. The differential LET spectrum, or Heinrich curve, gives the flux of all particles as a function of the linear energy transfer, and such curves can be derived for various orbits. The SEU rate, U, can then be calculated by integrating over the LET spectrum $f(L)$, and the path length distribution $p(l)$ in the sensitive volume, giving (ECSS, 2000 [19])

$$U = \frac{S}{4} \int_{E_c/L_{max}}^{l_{max}} p(l) \int_{E_c/l}^{L_{max}} f(L)\, dL\, dl \;, \tag{7}$$

where E_c is the critical (threshold) energy, L_{max} is the maximum LET expected, l_{max} is the maximum path length in the sensitive volume, and S is the total surface area of the sensitive volume. The lower limits in the integrals represent the shortest path capable of supporting upset (E_c/L_{max}) and the minimum particle LET necessary to cause upset on a path length l (E_c/l). The integration limits are established through testing. For the path length distribution, analytic expressions can be derived, if the shape of the sensitive volume is assumed to be a parallelepiped (Robinson et al., 1994 [89]). A survey of predictions and observations of SEU rates in space has been performed in Peterson (1997) [91].

5.5 Radiation-Induced Interference and Background in Instruments

Space missions are becoming steadily more demanding and the payload instruments more sophisticated with improved performance and sensitivity. Space radiation, and particularly the enhanced particle fluxes during various space weather events, are an increasing concern for instrument operation. Many examples exist of interference of energetic particles with scientific and technical spacecraft instruments. Due to the wide and frequent application of charge-coupled devices in current space missions, probably the most familiar effects presently are the high backgrounds in these sensors during large solar particle events. As an example, Fig. 9 presents energetic particle data and two CCD images from the famous July 14, 2000 solar event. The images are from the SOHO/LASCO coronagraph and the SOHO/EIT extreme-UV imaging telescope. The particle data are from the SOHO/ERNE particle instrument showing proton fluxes at 8 energy channels in the range 1.5–110 MeV. The backgrounds in the CCD images due to the particle-generated charge in a large number of pixels are very high and largely obscuring the targets, i.e., the solar corona (LASCO) and the sun itself (EIT).

Other reported examples of radiation-induced interference are the responsivity variations and "glitches" in the ISO infrared camera Isocam (Claret et

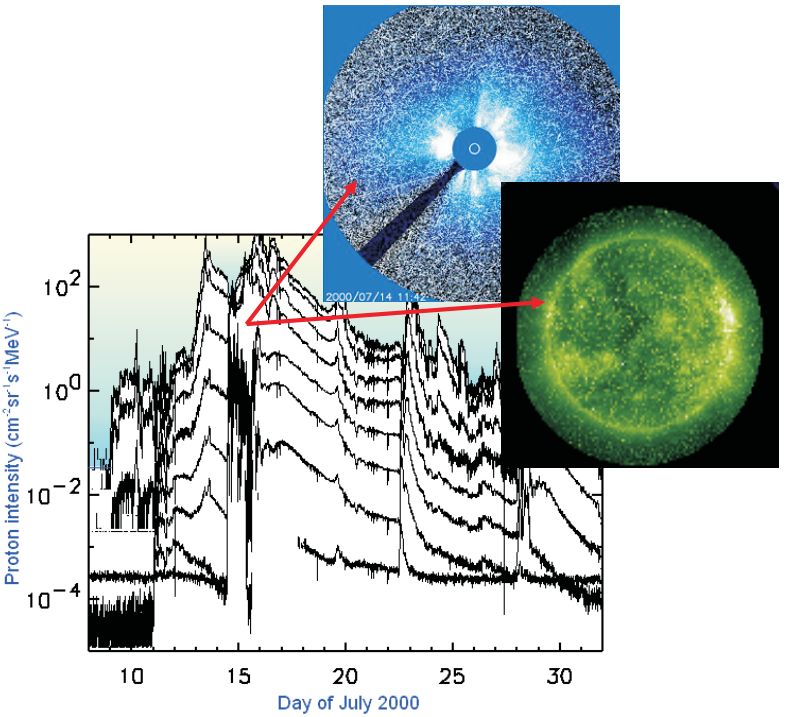


Fig. 9. Observed particle fluxes and particle-induced CCD background in July 14, 2000

al., 1999 [92]), and background signals in the Hipparcos telescopes (Daly et al., 1994 [93]). The Isocam response variations were found to be caused by the high radiation dose from trapped particles and related to the detector dark current, while the glitches were identified as transient effects due to various types of particles. The glitch occurrence was found to depend on space weather conditions. The background recordings in the Hipparcos Star Mapper were analysed in Daly et al. (1994) [93] and the results were applied in studying the dynamics of the radiation belts. The background was identified to be due to Cerenkov radiation and fluorescence produced by high-energy charged particles or by bremsstrahlung via Compton electrons passing through optical components, and due to secondary emission from photomultiplier window or electrodes induced by charged particle hits.

In general, sensors designed to sense photons also respond to energetic particles passing through them resulting in interference with scientific observations. Charged particle background is not a problem in photon detectors only, but as well in plasma and energetic particle detectors themselves. E.g., microchannel plate detectors and silicon particle detectors may have significant background count rates affecting their performance. In gamma ray de-

tectors (often various types of scintillators), secondary radiation generated by the primary environment near a sensor, including radioactive decay, can play a significant role in the sensor background.

The spurious background signals caused by energetic particles in various sensor systems not only deteriorate the quality of the recorded data. Particle-induced backgrounds also present complications in the form of decreased sensitivity, increased dead-time, and increased requirements for signal processing complexity (Vampola, 1989 [25]). Particle tracks in CCDs can also lead to inefficient performance of data compression algorithms (Kahler, 2001 [43]) and thereby decrease the rate of image transmission. Finally, it is worth mentioning that the mere thermal input by energetic particles to low-temperature sensors can be deleterious in some orbits. The transient energy input due to eneregtic particle populations can exceed $5\,\mathrm{W/m^2}$ (Vampola, 1989 [25]). Major geomagnetic storms producing large increases in energetic particle populations can increase the heat load averaged over an orbit by an order of magnitude, which can affect the performance of cryogenic sensor systems.

6 Summary and Final Remarks

A brief general overview of space weather effects on technology, and the space environment and its variations causing these effects, was given in this chapter. A number of space weather effects on spacecraft and their components were then discussed in some more detail. Space plasmas and particle radiation and their solar-induced enhancements are the basic sources of most of the functional and performance anomalies in technological systems encountered in space. Traditionally, spacecraft surface charging due to substorm injections of energetic plasma clouds has been regarded as the most important cause of environmental spacecraft anomalies, but several observations indicate that internal charging by very energetic electrons accelerated in connection of geomagnetic storms also play a significant role. Radiation effects discussed in this chapter included total ionising dose, displacement damage, and single event effects. There is a general trend of using more and more commercial off-the-shelf components in space. This also implies that TID effects become more and more important in future missions, and is amplified by the fact that total dose effects in commercial components strongly depend on the manufacturing process used, and even for the same manufacturer can vary from a manufacturing lot to an other. In addition, high-performance, very high-density microcircuits are more prone to various types of single event effects. Radiation effects due to displacement damage are emphasised in scientific missions, where instruments of many different scientific disciplines employ sensitive charge-coupled devices. Presently, as also demonstrated in this chapter, particle-induced background and interference in a hard radiation environment is a concern, but with increasing sensitivity requirements, displacement damage may become the limiting factor.

There is no doubt that space weather does affect systems in space. The two recent studies Harboe-Soerensen et al. (2002) [94] and Fieseler et al. (2002) [95] demonstrate the various effects in scientific satellites in two different environments. It is clear that pre-flight modelling of the environment and the response of systems in space to that environment are required. Sensitive science instruments also need detailed simulations to evaluate and minimise the effects, and to develop methods to remove the background caused by the environment. With the more demanding performance requirements and with the application of new technologies, the future systems are likely to be ever more vulnerable to space weather effects.

Acknowledgments

The author wishes to express his gratitude to the organizers of the International DPG Spring School on Space Weather for inviting him to give this lecture, and to the Deutsche Physikalischer Gesellshaft for the hospitality during his stay in Bad Honnef. Bernhard Fleck and Helmut Schweitzer, both at ESA-NASA/GSFC, are thanked for providing the SOHO solid state recorder and solar array data. SOHO is an international co-operation between ESA and NASA. The use of the CME catalog generated and maintained by the Center for Solar Physics and Space Weather, The Catholic University of America in cooperation with the Naval Research Laboratory and NASA is acknowledged. This work was partly supported by the Academy of Finland in the framework of the Antares programme.

References

1. The National Space Weather Program: The Implementation Plan, 2nd Edition, FCM-P31-2000 (2000) (http://www.ofcm.gov/nswp-ip/tableofcontents.htm)
2. M. Hapgood: 'Roadmap for European co-ordination in space weather', ESWS-RAL-RP-0003, Issue 1.0 (2001) (http://www.estec.esa.nl/wmwww/wma/spweather/esa_initiatives/spweatherstudies/public_doc.html)
3. S. Odenwald: Sky and Telescope, p. 50 (March 2000)
4. L.J. Lanzerotti, 'Space weather effects on technologies'. In: *Space Weather, Geophys. Monogr. Ser., vol. 125*, ed. by P. Song, H.J. Singer, G.L. Siscoe (AGU, Washington, D.C. 2001) pp. 11-22
5. H. Koskinen, E. Tanskanen, R. Pirjola, et al.: 'Space weather effects catalogue', ESWS-FMI-RP-0001, Issue 2.2 (2001) (http://www.estec.esa.nl/wmwww/wma/spweather/esa_initiatives/spweatherstudies/public_doc.html)
6. D.H. Boteler, R.J. Pirjola, H. Nevanlinna: Adv. Space Res. **22**, 17 (1998)

7. R. Pirjola, D. Boteler, A. Viljanen, O. Amm: Adv. Space Res. **26**, 5 (2000)
8. T.S. Molinski, W.E. Feero, B.L. Damsky: IEEE Spectrum, p. 55 (November 2000)
9. A. Pulkkinen, R. Pirjola, D. Boteler, A. Viljanen, I. Yegorov: J. Apl. Geophys. **48**, 233 (2001)
10. S. Basu, K.M. Groves, Su. Basu, P.J. Sultan: J. Atmosph. Solar-Terrestrial Phys. **64**, 1745 (2002)
11. S.H. Skone: J. Geodesy **75**, 457 (2001)
12. D.N. Baker: 'Satellite anomalies due to space storms'. In: *Space Storms and Space Weather Hazards, NATO Science Series II: Mathematics, Physics and Chemistry, vol. 38*, ed. by I.A. Daglis (Kluwer Academic Publishers, Dordrecht 2001) pp. 285-311
13. M.I. Panasyuk: 'Cosmic ray and radiation belt hazards for space missions'. In: *Space Storms and Space Weather Hazards, NATO Science Series II: Mathematics, Physics and Chemistry, vol. 38*, ed. by I.A. Daglis (Kluwer Academic Publishers, Dordrecht 2001) pp. 251-284
14. M.A. Shea, D.F. Smart: Adv. Space Res. **22**, 29 (1998)
15. D.H. Brautigam: J. Atmosph. Solar-Terrestrial Phys. **64**, 1709 (2002)
16. R. Gubby, J. Evans: J. Atmosph. Solar-Terrestrial Phys. **64**, 1723 (2002)
17. K.L. Bedingfield, R.D. Leach, M.B. Alexander: 'Spacecraft system failures and anomalies attributed to the natural space environment', NASA Reference Publication 1390 (1996) (http://trs.nis.nasa.gov/)
18. D.J. Gorney: Rev. Geophys. **28**, 315 (1990)
19. ECSS-E-10-04A, Space engineering: Space environment (ECSS Secretariat, ESA-ESTEC 2000) (http://www.ecss.nl/)
20. N.U. Crooker, G.L. Siscoe: 'The effect of the solar wind on the terrestrial environment'. In: *Physics of the Sun, Vol. 3*, ed. by P.A. Sturrock, T.E. Holzer, D.M. Mihalas, R.K. Ulrich (D. Reidel Publishing Company, Dordrecht 1986) pp. 193-249
21. H.V. Cane, G. Wibberenz, I.G. Richardson, T.T. von Rosenvinge: Geophys. Res. Lett. **26**, 565 (1999)
22. B. Klecker: Adv. Space Res. **17**, 37 (1996)
23. E.G. Stassinopoulos, J.P. Raymond: Proc. IEEE **76**, 1423 (1988)
24. R.A. Mewaldt, R.S. Selesnick, J.R. Cummings: 'Anomalous cosmic rays: The principal source of high energy heavy ions in the radiation belts'. In: *Radiation belts: Models and standards, Geophys. Monogr. Ser., vol 97*, ed. by J.F. Lemaire, D. Heynderickx, D.N. Baker, (AGU, Washington, D.C. 1996) pp. 35-41
25. A.L. Vampola: J. Spacecraft and Rockets **26**, 416 (1989)
26. E.J. Daly: Radiat. Phys. Chem. **43**, 1 (1994)
27. H. Fichtner: Space Sci. Rev. **95**, 639 (2001)
28. C.S. Dyer, P.R. Truscott, H. Evans, et al.: Adv. Space Res. **17**, 53 (1996)
29. J. Feynman, T.P. Armstrong, L. Dao-Gibner, S. Silverman: J. Spacecraft and Rockets **27**, 403 (1990)

30. J.W. Wilson, M-H.Y. Kim, J.L. Shinn, et al.: 'Solar cycle variation and application to the space radiation environment'; NASA/TP-1999-209369 (1999) (http://techreports.larc.nasa.gov/ltrs)
31. H. Friedman: Sun and Earth (Scientific American Books, Inc., New York 1986)
32. C.T. Russell: 'Solar wind and interplanetary magnetic field: A tutorial'. In: *Space Weather, Geophys. Monogr. Ser., vol. 125*, ed. by P. Song, H.J. Singer, G.L. Siscoe (AGU, Washington, D.C. 2001) pp. 73-89
33. A. Otto: The Magnetosphere, Lect. Notes Phys. **656**, 133–192 (2005)
34. A.C. Fraser-Smith: Rev. Geophys. **25**, 1 (1987)
35. M. Schüssler: The Sun and Its Restless Magnetic Field, Notes Phys. **656**, 23–49 (2005)
36. W.W. Vaughan, K.O. Niehuss, M.B. Alexander: 'Spacecraft environments interactions: Solar activity and effects on spacecraft', NASA Reference Publication 1396 (1996) (http://trs.nis.nasa.gov/)
37. J.T. Gosling: 'Coronal mass ejections: An overview'. In: *Coronal mass ejections, Geophys. Monogr. Ser., vol. 99*, ed. by N. Crooker, J.A. Joselyn, J. Feynman (AGU, Washington, D.C. 1997) pp. 9-16
38. J.A. Klimchuk: 'Theory of coronal mass ejections'. In: *Space Weather, Geophys. Monogr. Ser., vol. 125*, ed. by P. Song, H.J. Singer, G.L. Siscoe (AGU, Washington, D.C. 2001) pp. 143-157
39. O.C. St. Cyr, R.A. Howard, N.R. Sheeley, Jr., et al.: J. Geophys. Res. **105**, 18169 (2000)
40. L.F.E. Burlaga: 'Magnetic clouds'. In: *Physics of the inner heliosphere, Vol 2*, ed. by R. Schwenn, E. Marsch (Springer-Verlag, Berlin 1991) pp. 1-22
41. S.P. Plunkett, S.T. Wu: IEEE Trans. Plasma Sci. **28**, 1807 (2000)
42. D.V. Reames: Space Sci. Rev. **90**, 413 (1999)
43. S.W. Kahler: 'Origin and properties of solar energetic particles in space'. In: *Space Weather, Geophys. Monogr. Ser., vol. 125*, ed. by P. Song, H.J. Singer, G.L. Siscoe (AGU, Washington, D.C. 2001) pp. 109-122
44. B.T. Tsurutani, W.D. Gonzalez: 'The interplanetary causes of magnetic storms: A review'. In: *Magnetic storms, Geophys. Monogr. Ser., vol. 98*, ed. by B.T. Tsurutani, W.D. Gonzalez, Y. Kamide, J.K. Arballo (AGU, Washington, D.C. 1997) pp. 77-89
45. D.F. Webb, N.U. Crooker, S.P. Plunkett, O.C. St. Cyr: 'The solar sources of geoeffective structures'. In: *Space Weather, Geophys. Monogr. Ser., vol. 125*, ed. by P. Song, H.J. Singer, G.L. Siscoe (AGU, Washington, D.C. 2001) pp. 123-141
46. W.D. Gonzalez, B.T. Tsurutani: Planet. Space Sci. **35**, 1101 (1987)
47. H.V. Cane: 'The current status of our understanding of energetic particles, coronal mass ejections, and flares'. In: *Coronal mass ejections, Geophys. Monogr. Ser., vol. 99*, ed. by N. Crooker, J.A. Joselyn, J. Feynman (AGU, Washington, D.C. 1997) pp. 205-215
48. K.-L. Klein, G. Trottet: Space Sci. Rev. **95**, 215 (2001)

49. G.M. Mason, T.R. Sanderson: Space Sci. Rev. **89**, 77 (1999)
50. C.T. Russell: IEEE Trans. Plasma Sci. **28**, 1818 (2000)
51. D.N. Baker: J. Atmosph. Solar-Terrestrial Phys. **58**, 1509 (1996)
52. D.N. Baker, X. Li, N. Turner, et al.: J. Geophys. Res. **102**, 14141 (1997)
53. D.N. Baker, T.I. Pulkkinen, X. Li, et al.: J. Geophys. Res. **103**, 17279 (1998)
54. E.G. Mullen, M.S. Gussenhoven, K. Ray, M. Violet: IEEE Trans. Nucl. Sci. **38**, 1713 (1991)
55. M.S. Gussenhoven, E.G. Mullen, M.D. Violet: Adv. Space Res. **14**, 619 (1994)
56. M.S. Gussenhoven, E.G. Mullen, D.H. Brautigam: IEEE Trans. Nucl. Sci. **43**, 353 (1996)
57. D.N. Baker: J. Atmosph. Solar-Terrestrial Phys. **62**, 1669 (2000)
58. J.I. Vette: 'The AE-8 trapped electron model environment', NSSDC Report WDC-A-R&S 91-24, NASA-GSFC (1991)
59. D.M. Sawyer, J.I. Vette: 'AP-8 trapped proton environment for solar maximum and solar minimum', NSSDC Report WDC-A-R&S 76-06, NASA-GSFC (1976)
60. J. Feynman, G. Spitale, J. Wang, S. Gabriel: J. Geophys. Res. **98**, 13281 (1993)
61. J. Feynman, A. Ruzmaikin, V. Berdichevsky: J. Atmosph. Solar-Terrestrial Phys. **64**, 1679 (2002)
62. A.J. Tylka, J.H. Adams, Jr., P.R. Boberg, et al.: IEEE Trans. Nucl. Sci. **44**, 2150 (1997)
63. M.S. Gussenhoven, E.G. Mullen, D.H. Brautigam: 'Phillips laboratory space physics division radiation models'. In: *Radiation Belts: Models and standards, Geophys. Monogr. Ser., vol. 97*, ed. by J.F. Lemaire, D. Heynderickx, D.N. Baker (AGU, Washington, D.C. 1996) pp. 93-101
64. M.A. Xapsos, J.L. Barth, E.G. Stassinopoulos, E.A. Burke, G.B. Gee: 'Space environment effects: Model for emission of solar protons (ESP) – Cumulative and worst-case event fluences', NASA/TP-1999-209763 (1999) (http://trs.nis.nasa.gov/)
65. J. Barth: 'Modeling space radiation environments', 1997 IEEE NSREC Short Course Notes, Ch 1, Snowmass, CO, USA (1997) (http://radhome.gsfc.nasa.gov/ radhome/papers/SC_nsrec97.pdf)
66. J. Feynman, S.B. Gabriel: J. Geophys. Res. **105**, 10543 (2000)
67. H.C. Koons, J.E. Mazur, R.S. Selesnick, et al.: 'The impact of the space environment on space systems', Aerospace Report No. TR-99(1670)-1 (1999) (http://www.aero.org/publications/papers/tech-reports.html)
68. R.D. Leach, M.B. Alexander: 'Failures and anomalies attributed to spacecraft charging', NASA Reference Publication 1375 (1995) (http://trs.nis.nasa.gov/)
69. C.K. Purvis, H.B. Garrett, A.C. Whittlesey, N.J. Stevens: 'Design guidelines for assessing and controlling spacecraft charging effects', NASA Technical Paper 2361 (1984) (http://powerweb.grc.nasa.gov/pvsee/publications/thebasics.html)

70. E.C. Whipple: Rep. Prog. Phys. **44**, 1197 (1981)
71. R. Grard, K. Knott, A. Pedersen: Space Sci. Rev. **34**, 289 (1983)
72. H-L. Lam, J. Hruska: J. Spacecraft and Rockets **28**, 93 (1991)
73. J.J. Capart, J.J. Dumesnil: Esa Bull. No. 34, p. 22 (1983)
74. J. Goree, Y.T. Chiu: J. Spacecraft and Rockets **30**, 765 (1993)
75. G.L. Wrenn: J. Spacecraft and Rockets **32**, 514 (1995)
76. D.J. Rodgers, A.J. Coates, A.D. Johnstone, E.J. Daly: 'Correlation of Meteosat-3 anomalies with data from the space environment monitor', ESA WPP-155 (1998) (http://www.estec.esa.nl/ wmwww/wma/spweather/ workshops/proceedings_w1/ proceedings_w1.html)
77. A.R. Frederickson, E.G. Holeman, E.G. Mullen: IEEE Trans. Nucl. Sci. **39**, 1773 (1992)
78. F. Lei, P.R. Truscott, C.S. Dyer, et al.: IEEE Trans. Nucl. Sci. **49**, 2788 (2002)
79. A. Holmes-Siedle, L. Adams: Handbook of radiation effects, 2nd edition (Oxford Univerity Press, New York 2002)
80. J.R. Srour, J.M. McGarrity: Proc. IEEE **76**, 1443 (1988)
81. L.D. Edmonds, C.E. Barnes, L.Z. Scheick: 'An introduction to space radiation effects on microelectronics', JPL Publication 00-06 (2000) (http://nppp.jpl.nasa.gov/ resinfo_refmaterials.htm)
82. R.L. Pease: IEEE Trans. Nucl. Sci. **43**, 442 (1996)
83. T.R. Oldham, K.W. Bennett, J. Beaucour, et al.: IEEE trans. Nucl. Sci. **40**, 1820 (1993)
84. D. Bräunig, F. Wulf: Radiat. Phys. Chem. **43**, 105 (1994)
85. G.R. Hopkinson, Radiat. Phys. Chem. **43**, 79 (1994)
86. G.P. Summers, E.A. Burke, M.A. Xapsos: Rad. Measurements **24**, 1 (1995)
87. G.R. Hopkinson, C.J. Dale, P.W. Marshall: IEEE Trans. Nucl. Sci. **43**, 614 (1996)
88. N.V. Kuznetsov, R.A. Nymmik: Radiat. Measurements **26**, 959 (1996)
89. P. Robinson, W. Lee, R. Aguero, S. Gabriel: J. Spacecraft and Rockets **31**, 166 (1994)
90. W. Heinrich: Radiat. Phys. Chem. **43**, 19 (1994)
91. E.L. Petersen: IEEE Trans. Nucl. Sci. **44**, 2174 (1997)
92. A. Claret, H. Dzitko, J.J. Engelmann: IEEE Trans. Nucl. Sci. **46**, 1511 (1999)
93. E.J. Daly, F. van Leeuwen, H.D.R. Evans, M.A.C. Perryman: IEEE Trans. Nucl. Sci. **41**, 2376 (1994)
94. R. Harboe-Soerensen, E. Daly, F. Teston, et al.: IEEE Trans. Nucl. Sci. **49**, 1345 (2002)
95. P.D. Fieseler, S.M. Ardalan, A.R. Frederickson: IEEE Trans. Nucl. Sci. **49**, 2739 (2002)

Radiation Risks from Space

Juergen Kiefer

University Giessen, Germany

Abstract. Radiation from space has many influences on human life, of course in space but also on Earth. This chapter describes the particular structure of the radiation field which consists - contrary to the terrestrial situation - mainly of charged nuclei from protons to nickel nuclei. Their interaction with the Earth's environment is discussed. Understanding the health risks originating from space radiation presents a challenge to biophysical research, some of the current approaches are outlined. Quantitative risk assessment for humans is still very difficult as epidemiological data are not available for the action of charged particles. The concepts used are based on radiation protection systematics which are described.

1 Introduction

The history of space radiation research started in 1912 when the Austrian physicist Viktor Franz Hess (1883-1964) sent detectors into the upper atmosphere by using balloons to measure the attenuation of terrestrial radiation with distance. Quite to his surprise he found an increase in radiation intensity rather than the expected reduction. He interpreted this by postulating what we nowadays call "space radiation" and later he could also show that the main source is not the sun but the galaxy. Deservedly Hess received the Nobel Prize in 1936 for his work. Thanks to the modern developments which began with the famous "sputnik" in 1957 we are now in the possession of a wealth of data although the origin of galactic cosmic rays remains still an enigma.

The radiation field in space may pose a considerable hazard to astronauts but it extends its action also to the terrestrial environment. Its composition is quite different from that on earth, it contains mainly charged particles of considerable energies. The analysis of their biological effects confronts the researcher with new and demanding problems, both theoretically and experimentally. This chapter will summarise some of the results obtained. It starts with the radiation sources and then discusses the important biological results which will be used to delineate the possible risks involved.

J. Kiefer, Radiation Risks from Space, Lect. Notes Phys. **656**, 275–292 (2005)
http://www.springerlink.com/

2 Radiation Sources

2.1 The Extraterrestrial Field

As far as effects on humans are concerned only the particle component of the space radiation has to be taken into account. Generally speaking it consists of 85% protons, 14% helium ions (alpha particles) and to about 1% of heavier nuclei, extending as far as nickel. Their relative abundance is shown in Fig. 1, the total fluence rate is about $4\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$ at solar minimum (see below). The energies span over a wide range with a maximum of about 10^7 MeV/u. The differential fluence distributions are qualitatively similar for all ions peaking between 100 and 1000 MeV/u. These particles possess considerable ranges (see Fig. 2 for protons as example) and can easily penetrate spacecraft walls as well as the human body. They interact with shielding material giving rise to secondary radiations via fragmentation processes and nuclear reactions. In terms of biological significance they may be more relevant than the primaries, particularly neutrons play an important role in this respect.

Also the sun's emissions add to the radiation environment in outer space, mainly also by protons and alpha particles. Their energies are lower than those of the "Galactic Cosmic Rays" (GCR) but their intensities may rise to dangerous levels during "solar particle events" (SPE, see below). The sun influences also the intensity of the GCR component via magnetic interactions resulting in a reduction by about one order of magnitude at solar maximum (Fig. 3).

2.2 Trapped Radiation: The Radiation Belts

The space charged particles interact with the earth's magnetic field and are captured if their rigidity (i.e. momentum per charge) is below a certain limit. This leads to the formation of "radiation belts" (also called "van Allen belts", named after their discoverer who found them during the early "explorer"

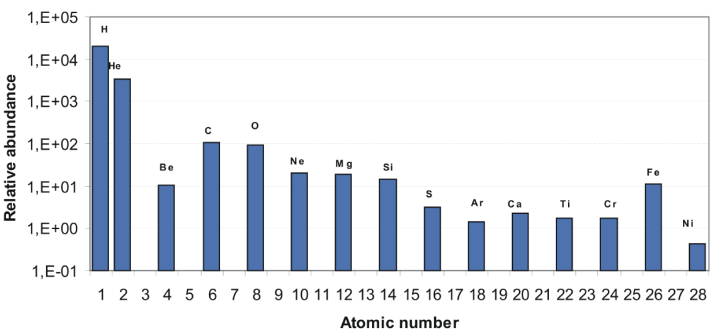


Fig. 1. The relative abundance of charged nuclei in cosmic galactic radiation (Simpson, 1983 [13])

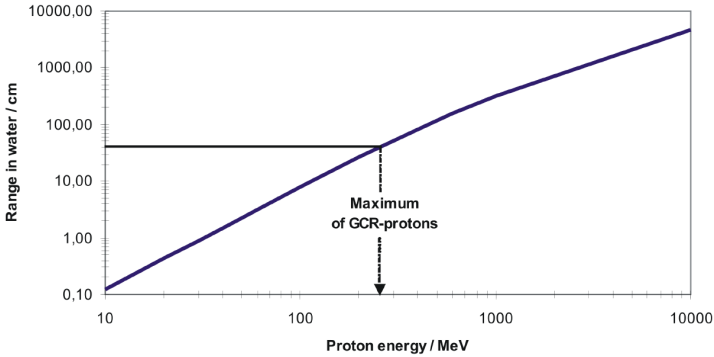


Fig. 2. Ranges of protons in water (equivalent to soft tissue) as a function of energy

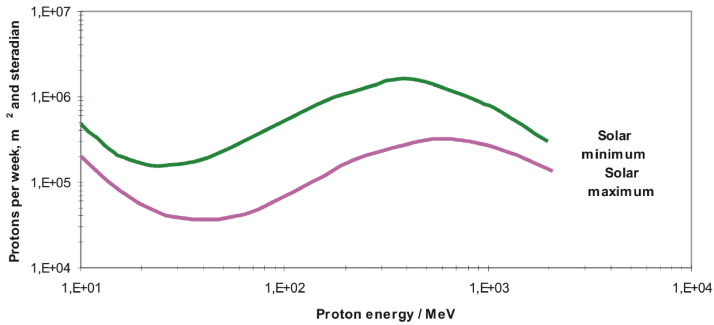


Fig. 3. The modulation of GCR by the solar field (Silberberg et al., 1984 [12])

missions in the 1960s). The particles gyrate along the magnetic field lines and are reflected at the maximum field strength. There is a strong dependence on altitude and latitude. Near the poles of the earth magnetic field the radiation belts reach down to fairly low altitudes. Because of their charge electrons and protons drift in opposite directions, electrons eastward, protons and heavier nuclei westward.

There are two radiation belts for electrons extending to about 2.4 earth radii and between 2.8 to about 12 earth radii over the equator, respectively. The two zones differ in electron energy, the lower contains particles with less than 5 MeV, while the spectrum in the outer one is much harder (around 7 MeV). Because of the low penetrating ability trapped electrons do not constitute a significant hazard to the inner of satellites in low earth orbits (LEO) but they may damage surfaces (e.g. of open solar panels) and have to be taken into account with extravehicular activities (EVAs) where only comparatively thin space suits are worn. The surface dose rates are in the order of Grays per day which were measured during the passage through the South Atlantic Anomaly (SAA, see below).

For protons there are no clearly defined belts. Also their energy increases about inversely with the distance to the earth, i.e. the most energetic protons are found close to the earth. Only few heavier ions are trapped and they have low energies, so that they have not to be taken into account for risk analysis. They are, however, an important component of GCR as said above.

The earth magnetic field is not symmetrical but distorted in certain areas. The most important in the present context is the "South Atlantic Anomaly" (SAA), westward of the South American continent where the magnetic field strength is significantly reduced so that the radiation belt reaches down deep into the atmosphere and to altitudes of low earth orbits. The highest doses are recorded when satellites cross the SAA which are mainly due to protons and secondary radiations produced by their interaction with spacecraft materials.

2.3 Interactions of Space Radiations with the Atmosphere

GCR particles do not reach the earth surface. They interact with the molecules in the upper atmosphere producing secondary radiations. The most important are μ-mesons (myons), electrons and neutrons. Their fluences depend on altitude: neutrons are strongly absorbed on their way downwards and do not play any significant role at sea level. The most important contribution to environmental natural radiation from cosmic radiation is due to myons and electrons. This relationship changes with height, at flight altitudes neutrons contribute quite significantly as shown in Fig. 4 (see also Sect. 4).

Cosmic radiation adds also in another way to the terrestrial radiation budget, namely by the generation of cosmogenic radionuclides, which are formed in the upper atmosphere by nuclear reactions. The most important are:

$$^{14}\mathrm{N} + n \quad \rightarrow \quad ^{14}\mathrm{C} + p \;\; \text{(Radiocarbon)} \tag{1}$$

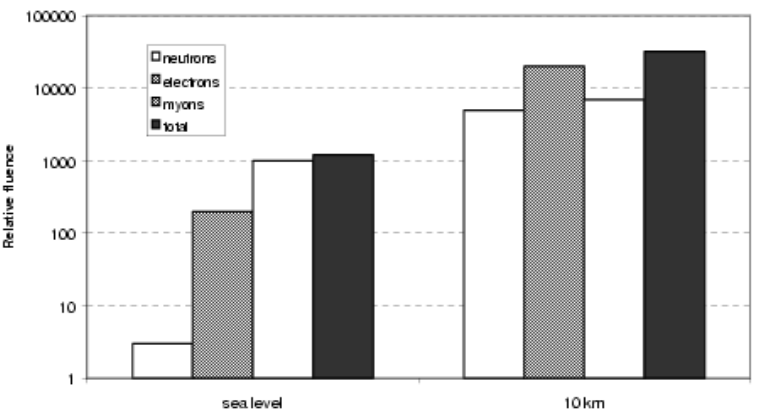


Fig. 4. The change of the contribution of secondary particles with altitude (Kiefer and Kölzer, 1986 [3])

and

$$^{14}\mathrm{N} + n \quad \rightarrow \quad ^{3}\mathrm{H} + ^{12}\mathrm{C} \ \ (\text{Tritium}) \tag{2}$$

Others of some importance are ^{7}Be and ^{22}Na. These radionuclides enter the food chain and are ingested by humans adding to the natural radiation burden although only to a small extent (Sect. 4).

2.4 Solar Particle Events (SPE)

Under normal circumstances the contribution of solar particle radiations is comparatively small but this situation may change rather dramatically in the case of sudden eruptions. The particle fluences during these "solar particle events" (SPE) may increase by order of magnitudes creating serious hazards to astronauts and electronic equipment. SPEs occur mainly during the periods of high solar activity and are rare during solar minimum. They are virtually unpredictable but they may be early detected as they are accompanied by X-ray bursts which reach monitoring instruments earlier than the charged particles. This gives a certain time span for warning.

Figure 5 gives an example of proton fluences recorded during one of the largest events in 1972. Although the energies of the most abundant protons – and hence the penetration power – is small compared to GCR-protons the increase by about four orders of magnitude in the 100 MeV-range is still very substantial constituting large problems for shielding and shelter of astronauts. This is a particularly serious in the case of long-term flights, e.g. to Mars.

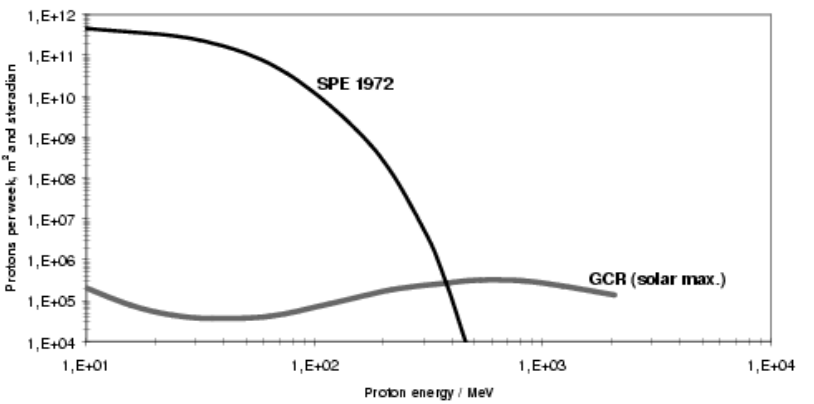


Fig. 5. Proton fluences during the very large solar particle event in 1972 compared to GCR-protons at solar maximum (Silberberg et al., 1984 [12])

3 The Biophysics of Space Particles

3.1 Energy Deposition by Charged Particles

Contrary to photon radiations and swift electrons ("sparsely ionising radiations") heavier charged particles deposit energy in densely ionising tracks where the ionisations lie close to the ion's trajectories. To describe this in a quantitative manner the quantity "linear energy transfer" (LET) has been introduced. It is defined as the energy locally imparted per unit pathlength. The common unit is "keV/μm". LET has a close relationship to the more familiar "mass stopping power", this is obtained by dividing LET by the density of the interacting medium. There is, however, a difference in philosophy: While stopping power deals essentially with the energy loss of the incoming particle, LET concentrates on the energy transferred to the exposed matter. The term "locally" has an important meaning in this context as discussed below.

LET depends both on particle effective charge (z_{eff}) and its energy, it increases with charge and decreases essentially inversely with energy. When the ion slows down it picks up electrons, and the effective charge is reduced so that LET decreases again at very low energies. This means that LET passes a maximum at low energies as shown for protons in Fig. 6. At equal energies the linear transfer scales essentially with z_{eff}, examples for some important space particles are shown in Fig. 7.

The most important quantity to describe radiation interaction with biological material is the "dose", defined as the energy absorbed per unit mass. Its dimension is J/kg which has been given the special name "Gray" (Gy) named after the English radiation physicist and biologist Louis Harold Gray.

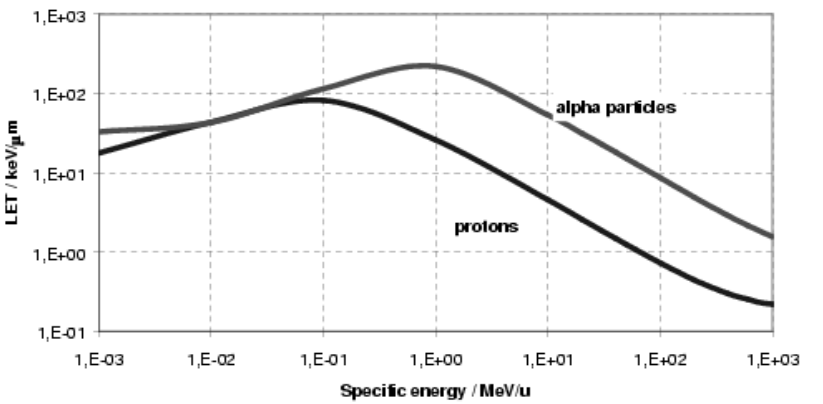


Fig. 6. Linear energy transfer of protons and alpha particles as function of particle specific energy

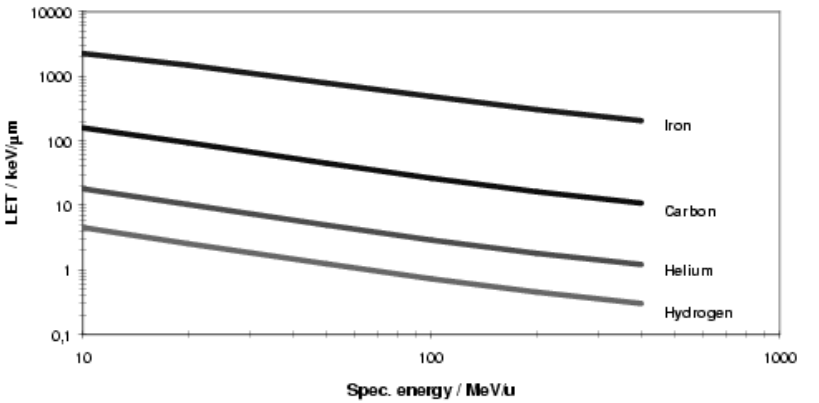


Fig. 7. Linear energy transfer as a function of particle specific energy for various ions

Dose and particle fluence are linked through the linear energy transfer:

$$D = \frac{\Phi L}{\rho} \tag{3}$$

where D stands for dose, L for LET and ρ for the density of the absorbing medium.

LET is quite low for very energetic particles, particularly with protons as demonstrated in Fig. 7, with other words they do not deliver large doses. This statement, however, holds only at the entrance surface. When they are slowed down on their way through matter they lose energy whereby LET increases. Just before they reach the end of their range there is a steep rise in local energy deposition (Fig. 8) which is called the "Bragg peak". The situation is not only important in terms of dose but also for the biological effect as the effectiveness depends on LET as discussed in Sect. 3.3 below.

The expression "local" in the definition of LET creates some difficulties. Charged particles interact via ionisations thus liberating electrons. Depending on the speed of the incoming ion they may have considerable energies and thus be able to travel large distances from the point of their origin. Within the framework of LET this is accounted for by introducing "restricted" LET where only electrons below a certain cut-off energy are counted indicated by an index. Customary is a cut-off value of 100 eV (LET_{100}) corresponding to a range of about 5 nm in water. LET_{∞} (unrestricted LET) where all electrons are counted is thus numerically equal to the stopping power.

3.2 Track Structure

An ion's path through matter does not resemble a "pencil beam" as just indicated but rather a "test tube brush" where the bristles are made up of the

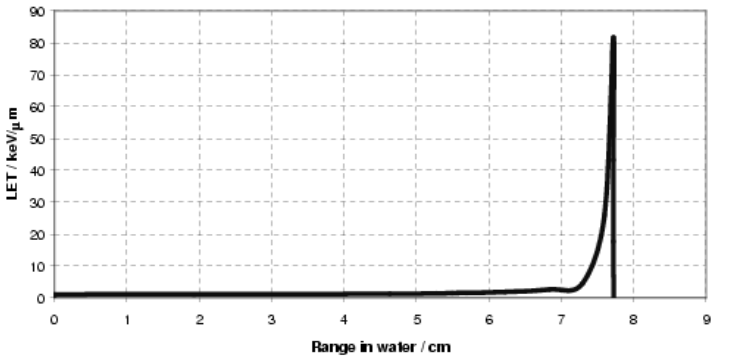


Fig. 8. Linear energy transfer of protons versus range in water.

secondary electrons. This track structure has to be taken into account by any attempt to understand the action of heavy charged particles in a quantitative way. With protons and alpha particles it is now possible to calculate the microscopic pattern of energy deposition by Monte Carlo computations but not yet for heavier ions as the interaction cross are not sufficiently known. Another approach, pioneered by R. Katz and sometimes named "amorphous track model" is to consider "local doses" within the track, i.e. the average energy in cylindrical shells around the ion path. As emission angel and electron energy are related there is a maximum radial extension of the track which is called the "penumbra radius" r_p. According to a current model (Kiefer and Straaten, 1986 [4]) the following relation holds

$$r_p = 0.0616E^{1.7} \tag{4}$$

where r_p is measured in μm and E is the ion specific energy in MeV/u. The penumbra radius does not depend on ion charge but local doses in the track do.

For a given energy the electron density scales with z^2_{eff} and the local dose $d(r)$ is given by the following formula (Kiefer and Straaten, 1986 [4])

$$d(r) = 1.25 \times 10^{-4} \frac{z^2_{\text{eff}}}{\beta^2 r^2} \tag{5}$$

with $\beta = v/c$ being the ion speed relative to that of light *in vacuo*, r the radial distance from the track centre in mm, $d(r)$ is given in Gy.

Figure 9 displays an example for 10 MeV/u iron ions. One sees that close to the track centre very high local doses are achieved. This plays an important role for the understanding of the increased biological effectiveness of heavy charged articles because very complex lesions can be created.

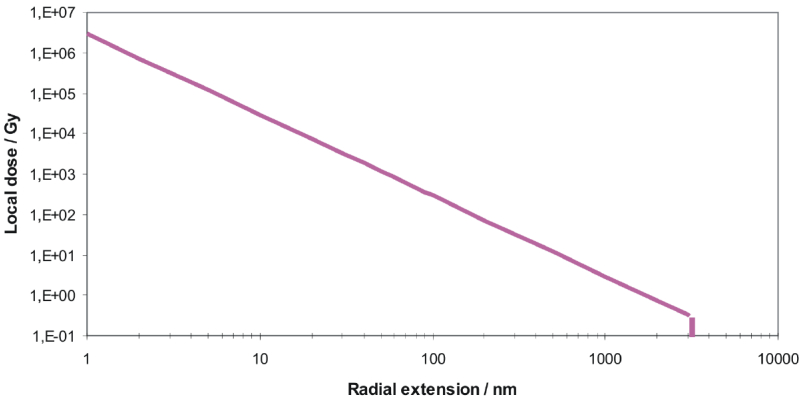


Fig. 9. The distribution of local dose in the track of a 10 MeV/u iron nucleus after Kiefer and Straaten (1986) [4].

3.3 Examples of Biological Results. The Difference Between Sparsely and Densely Ionising Radiations

Mutations play an essential part in the genesis of tumours and may also lead to hereditary defects in the progeny of radiation exposed people. They are, therefore, chosen as an example. Figure 10 shows mutation induction in mammalian (V79 Chinese hamster) cells after exposure to Xrays and ^{241}Am-α-particles which have under the experimental conditions a specific energy of about 0.75 MeV/u. Details are omitted here, they can be found in the

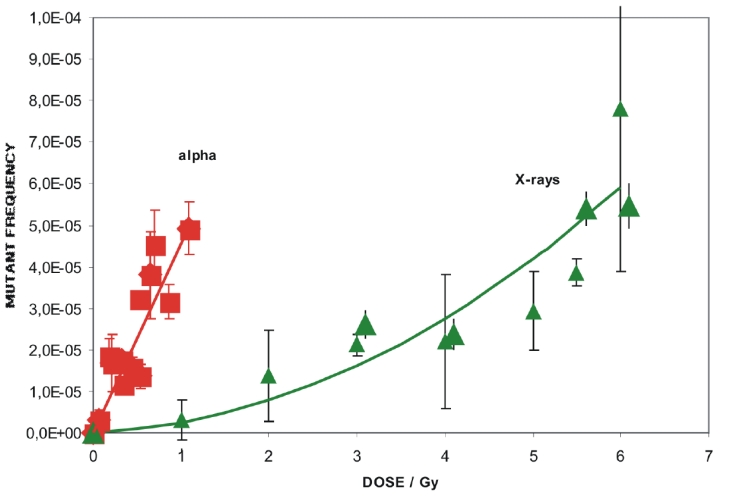


Fig. 10. Mutant frequencies in Chinese hamster cells as a function of ^{241}Am-α-particle and X-ray dose (Schmidt and Kiefer, 1998 [11])

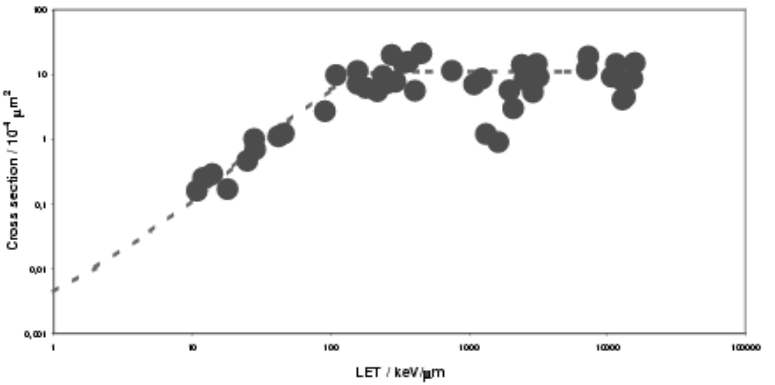


Fig. 11. Mutation induction cross section versus LET (Kiefer et al., 2001 [5]).

original publication (Schmidt and Kiefer, 1998 [11]). One notes immediately that at equal doses the effects are strikingly different, also the shape of the dose-effect curve changes with radiation quality. Obviously alpha particles are considerably more effective than X-rays. A quantitative measure of this important difference is the quantity "Relative Biological Effectiveness" (RBE for short) which is defined as the ratio of iso-effect doses. It is clear from the diagram that RBE is not constant with dose because of the differences in the shapes of the curve. The highest values are reached at very low doses where normally experimental data are difficult to obtain and have large errors. This constitutes an eminent problem in radioprotection which cannot be further elaborated here. RBE is related to the weighting of different radiation types in radiation protection as further discussed in the next section.

Mutant frequencies, i.e. the fraction of mutated surviving cells, are often plotted versus ion fluence which is related to dose according to (3). The slope of the linear relationship (which is obtained in most cases, see Fig. 10 for alpha particles) has thus the dimension of an area and is commonly termed "mutation induction cross section". This follows common use in physics. It has no geometrical meaning but gives only the probability of one particle to cause an effect. There is quite some scatter of the data which is unfortunately the rule with this kind of experiments illuminating the variability of biological responses.

The induction cross section is a useful parameter to describe the efficiency of different ions to cause biological effects. We have studied a large number of ions, mainly generated by the heavy ion accelerators at the "Gesellschaft für Schwerionenforschung" (GSI) at Darmstadt, Germany (Stoll et al., 1995 [14], 1996) with different biological endpoints, the mutation data were summarised recently (Kiefer et al., 2001 [5]). Figure 11 displays mutation cross sections versus LET_∞ for some selected particles. They rise initially and reach a plateau beyond about 100 keV/μm. This behaviour reflects the fact that ions

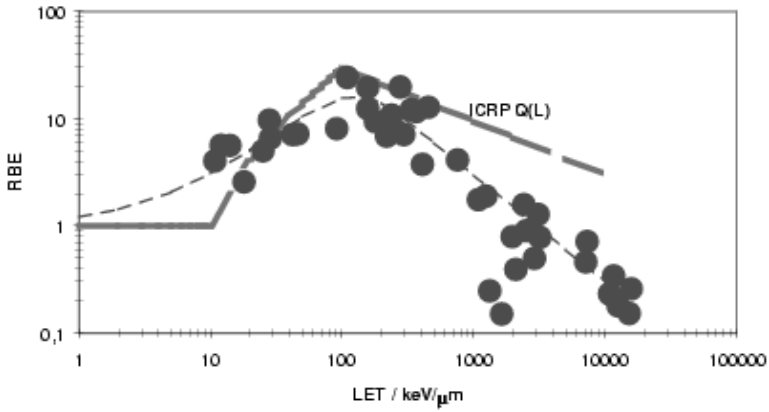


Fig. 12. Relative biological effectiveness for mutation induction by heavy ions related to X-rays. For comparison the ICRP $Q(L)$ relationship is also shown.

with higher ionisation density are more potent then the lighter species. The broken line is a semi-empirical fit to the data. There is no further increase in the high LET-region indicating that the efficiency per ion saturates.

Cross sections can also be used to calculate RBE-values. They are shown for several ions in Fig. 12. One notes that a maximum is found around an LET of 100 keV/μm. For comparison the $Q(L)$ relationship as suggested by ICRP is also shown (see below).

4 Approaches to Risk Assessment

4.1 Approaches and Quantities in Radiation Protection

Radiation protection is both a scientific discipline and a (practical) system to protect people from the harmful effects of radiation which is laid down in legal procedures. In order to keep it manageable a certain degree of simplification is unavoidable which leads necessarily to some lack of scientific scrutiny. One has, therefore, created special quantities only to be used in radiation protection. This has to be stressed as one finds quite often some confusion, not only in the general public but even in the scientific community.

The overriding aim of radiation protection is to reduce radiation related risk to people who are exposed during their professional activities. Table 1 gives a simplified overview about health hazards which can be related to ionising radiations. Acute effects develop – depending on dose – in short times, mostly in organs with high cell turnover and may even lead to acute death. They are in principle predictable in the individual case and occur only above certain threshold doses. Under normal circumstances they do not pose a problem as threshold doses are considerably higher than encountered at the working place. This is even true for astronauts but not in the case of

Table 1. Radiation health hazards and their general properties

Acute effects:	**Late effects:**
Malfunction of organs, eye cataracts, embryonic misdevelopment	Genetic risk by germ cell mutations, cancer
Deterministic, *Threshold doses*	*Stochastic,* *No threshold doses*

Table 2. Radiation weighting factors w_R for different types of ionising radiations (ICRP, 1991 [2])

Radiation type	w_R
photons, electrons	1
neutrons (depending on energy)	5 - 20
protons	5
alpha particles, heavy nuclei	20

large solar particle events where acute effects might even impair the crew performance. Late effects, on the other hand, constitute the real problem. Threshold doses are not known so that even small doses may pose a risk although with low, but not vanishing, probability. Genetic damage which may be transmitted to future generations has to be considered for younger people, the main concern is, however, cancer. There is a solid although not sufficient body of epidemiological data to delineate the human radiation related cancer risk (see e.g. UNSCEAR, 2000 [19] for the most recent review). The still most important study is the follow-up of the fate of the Japanese atomic bomb survivors which forms the basis also of radiation protection measures (see Preston et al., 2003 [9] for a most recent summary).

It has been shown above that different types of radiation possess different effectivities to lead to biological actions as described by the term RBE. To account for this in radioprotection the new quantity "equivalent dose" has been introduced which is obtained by multiplying the physical dose (measured in Gray) by "radiation weighting factors" w_R (Table 2). The unit of this new quantity is "Sievert" (abbreviated Sv). Although nearly every organ of the human body may develop cancer as a consequence of radiation exposure there are distinct differences. To account for this the most important organs or tissues are given "tissue weighting factors" w_T, listed in Table 3. The large value for the gonads does not reflect their high sensitivity in terms of cancer risk (which is in fact very low for males and small for females) but takes the genetic risk into account. The practically most important quantity, the "effective dose" is obtained by multiplying the organ equivalent doses by the

Table 3. Organ weighting factors w_T (ICRP, 1991 [2])

Organ or tissue	w_T	Organ or tissue	w_T
gonads	0.20	liver	0.05
red bone marrow	0.12	breast	0.05
colon	0.12	oesophagus	0.05
lung	0.12	thyroid	0.05
stomach	0.12	skin	0.01
bladder	0.05	bone surfaces	0.01
		remainder	0.05

Table 4. Quantities and units used in radiation protection

Quantity	Definition	Unit	name
Dose:	Energy imparted per unit mass	J/kg	Gray (Gy)
Dose equivalent:	Dose × (radiation) quality factor	J/kg	Sievert (Sv)
Effective dose:	$\sum$ [Organ (equivalent) dose × organ weighting factor]	J/kg	Sievert (Sv)

applicable organ weighting factor and summing up over all organs (Table 4). Effective doses are also measured in Sievert.

The system of radiation weighting factors is not suitable for the space situation as it neglects differences between ions and the influence of particle energy. In this case w_R can be replaced by the "quality factor" $Q(L)$ which is also used for operational quantities like e.g. ambient dose. The suggested functional relationship (ICRP, 1991 [2]) has the form (L stands for LET)

$$\begin{aligned} Q(L) &= 1 && \text{for } L < 10\text{keV}/\mu\text{m}, \\ Q(L) &= 0.32L - 2.2 && \text{for } 10 \le L \le 100\text{keV}/\mu\text{m}, \\ Q(L) &= 300\sqrt{L} && \text{for } L > 100\text{keV}/\mu\text{m} \end{aligned}$$

The "International Commission on Radiological Protection" (ICRP) derived "nominal risk factors" for fatal risk which are summarised in Table 5 (ICRP, 1991 [2]). It has to be pointed out that the values given are "nominal" in the sense that they neglect differences in age and gender or ethnic background and constitute thus "best estimates" for the derivation of dose limits, not intended and not suitable to calculate expected death rates as a result of radiation exposure as it is sometimes done in an uncritical and erroneous way.

In order to protect radiation workers from unacceptable high risk limits of effective doses are introduced which are set by legislation in most countries. They cover the danger of late effects, to avoid acute effects there are also dose limits for some organs. Values are given in Table 6, together with recommendations for space which are discussed below.

Table 5. Nominal risk factors for stochastic diseases in %/Sv (ICRP, 1991 [2])

Organ or tissue	nom. risk	Organ or tissue	nom. risk
Cancer:			
		oesophagus	0.30
bladder	0.30	ovaries	0.10
bone marrow	0.50	skin	0.02
bone surfaces	0.05	stomach	1.10
breast	0.20	thyroid	0.08
colon	0.85	remainder	0.50
liver	0.15	**total** (cancer)	5.00
lung	0.85		
		severe heritable disorders	1

Table 6. Yearly dose limits in Sv for occupationally exposed people on Earth (Strahlenschutzverordnung, 2001 [16]) and in space (NCRP, 2001 [8])

Organ	Earth (Germany 2001)	Space NCRP 132 (2000)
Eye	0.15	2
Skin	0.5	3
Bone marrow	0.05	0.5
Effective dose	0.02	0.5

4.2 Radiation Risks from and in Space

Humans on Earth are exposed to a number of radiations, both natural and by technical or medical activities. Figure 13 gives an overview for the average citizen in Germany from which it is seen that cosmic radiation gives only a comparatively small contribution, namely 0.3 mSv per year at ground level,

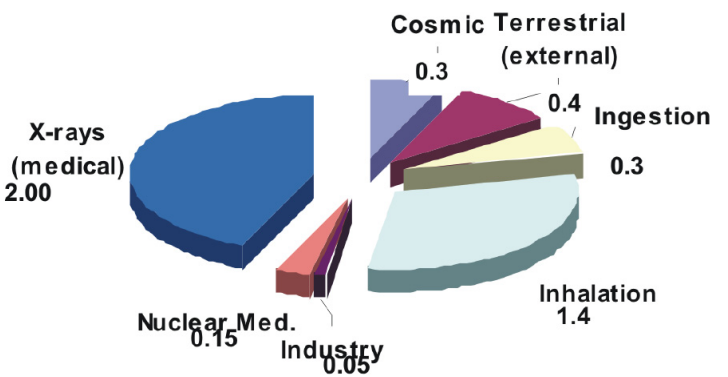


Fig. 13. The average radiation burden (mSv/year) for German citizens (Values for 1997, Bundesamt für Strahlenschutz)

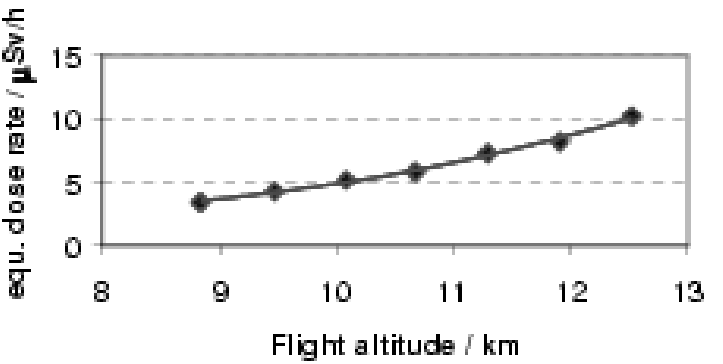


Fig. 14. Equivalent dose rates at different altitudes (Regulla and David, 1993 [10])

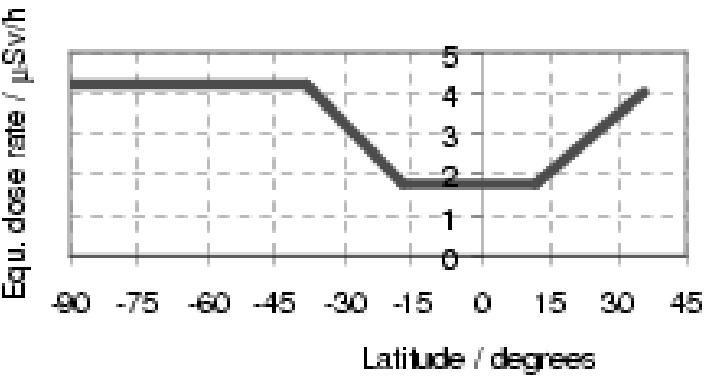


Fig. 15. Dependence of flight doses on latitude (Regulla and David, 1993 [10])

mainly due to the charged secondary particles, neutrons are negligible at see level. The situation is quite different at flight altitudes: Not only the doses are higher but also the composition of the radiation field changes as shown in Fig. 4 above. As in the recent legislation flight personnel has to be considered as "professionally exposed" quite a number of measurements have been performed for typical air routes. Figure 14 gives a summary which shows that for typical flight altitudes between 10 and 11.7cm km equivalent dose rates of 5-8 μSv/hour may be expected. These values refer to geomagnetic latitudes above 50° where the radiation belts exert some influence, they are considerably lower near the equator as depicted in Fig. 15. Taking into account the actual flight configuration one may estimate that the total effective dose to a passenger travelling from Europe to USA amounts to about 50–65 μSv which corresponds roughly to that of a thorax X-ray or less than 3% of the yearly dose average from natural environmental sources. It is well below any threshold for deterministic health effects, also clearly of that for teratogenic actions in the growing embryo.

Table 7. Annual effective doses caused by cosmogenic!radionuclides. Doses are in μSv (UNSCEAR, 1993 [18])

Nuclide	annual dose μSv
^{3}H	0.01
^{7}Be	3
^{14}C	12
^{22}Na	0.2

Cosmogenic radionuclides enter the food chain and may thus be ingested. The resultant doses are very low indeed as summarised in Table 7, they are much lower than the contribution of the primordial isotope ^{40}K.

The radiation situation in space (in satellites, the space station or on voyages to the moon or to other planets) is considerably more serious. Space travelling is, of course, always risky but virtually all possible hazards are restricted to the actual flight – except for a radiation sequel. The danger of acute effects exists only in the case of large solar particle events, here it may be, however, quite real as seen in Fig. 16 where the total doses for the entire event (a few hours duration) are compared with terrestrial and space limits. How to deal with unforeseen (and unpredictable) SPEs is still an unsolved problem with long-terms flight in outer space, e.g. for a Mars mission.

The contribution of galactic cosmic and solar radiation is less dramatic but the doses to be expected on the shuttle or the space station are still considerably higher than on Earth. Figure 17 shows some best estimates based on many measurements. The differences between solar minimum (1977) and solar maximum (1970) are obvious. The values were calculated with different shielding masses of polythene which is to be preferred as it reduces secondary

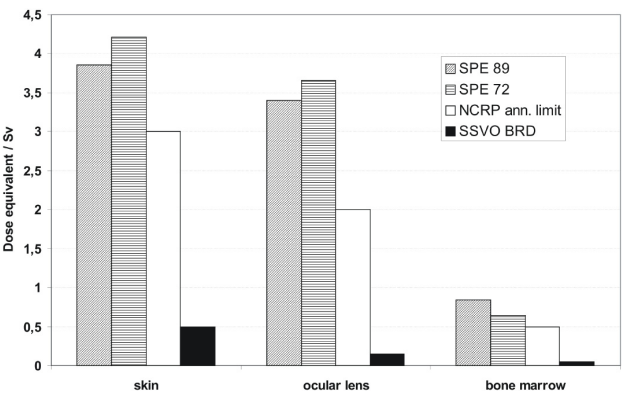


Fig. 16. Organ doses during large solar particle events in 1972 and 1989 (Townsend et al., 1991 [17]) compared to dose limits in space (NCRP, 2001 [8]) and on Earth (Strahlenschutzverordnung, 2001 [16])

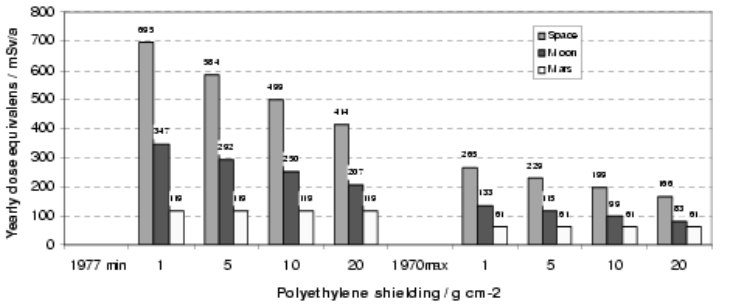


Fig. 17. Effective doses estimated for open space, the moon and Mars due to galactic cosmic radiation with different polyethylene shieldings and for solar minimum (1977) and solar maximum (1970) (G. Reitz, personal communication)

radiations more effectively (Wilson, 2000 [20]). With equal masses of aluminium the doses are still higher. Mars voyage scenarios presently discussed foresee duration between about 400 and 950 days. Even if performed during maximum solar activity (with a higher probability of solar flares which are not included in the estimates given) the expected doses approach a value of 1 Sv which is 2.5 times the "career limit" allowed according to German legislation.

5 Concluding Remarks

Radiation from space influences humans in many respects as discussed, both on Earth and in space. We are still far from understanding the risk of cosmic rays and solar particle events. Although an impressive body of measurements is available there is still an element of uncertainty, particularly with solar particle events. The lack of knowledge is still more pronounced with regard to the biological effects (see Cucinotta et al., 2002 [1] for an extensive discussion). The problems concerned are very interesting for biophysicists, giving them fascinating problems at the borderline of physics, biology and medicine.

References

1. Cucinotta, F., Badhwar, G., Sagani, Schimmerling, W., Wilson, J., Peterson, L., and Dicello, J., Space radiation cancer risk projections for exploration missions: Uncertainty reduction and mitigation, NASA Technical paper TP-2002–210777, 2002.
2. ICRP 60 ICRP International Commission on Radiological Protection 1990 recommendations of the ICRP, Ann. ICRP 21 No. 1–3, 1991.
3. Kiefer, H., and Kölzer W., *Strahlen und Strahlenschutz*, Springer-Verlag Berlin-Heidelberg-New York-Tokyo, 1986.

4. Kiefer, J., Straaten, H. 1986. Physics in Medicine and Biology 31, 1201-1209.
5. Kiefer, J., Schmidt, P., and Koch, S., Radiat. Res. 156, 607–611, 2001.
6. NCRP National Council on Radiation Protection and Measurements, Guidance on radiation received in space, Report 98, Bethesda Md., 1989.
7. NCRP National Council on Radiation Protection and Measurements, Radiation protection guidance for activities in low-earth orbit, Report 132, Bethesda Md., 2000.
8. NCRP National Council on Radiation Protection and Measurements, Bethesda Md., 2001.
9. Preston, D.L., Shimizu, Y., Pierce, D.A., Suyama, A., and Mabuchi, K., Radiat. Res. 160, 381–407, 2003.
10. Regulla, D., and David, J., Radiation measurements in civil aviation, GSF-Report 41/91 (English translation), GSF Neuherberg, 1993.
11. Schmidt, P. and Kiefer, J., Mutat. Res. 421, 149–161, 1998.
12. Silberberg, R., Tsao, C. H., Adams, J. H., and Letaw, J. R., Adv. Space Res. 4, 143–149, 1984.
13. Simpson, J. A., *Introduction to the galactic cosmic radiation* in: Composition and origin of cosmic rays, ed.: Shapiro, M.M., Reidel Publ. Dordrecht, Netherlands, 1983.
14. Stoll, U., Schmidt, A., Schneider, E., and Kiefer, J., Radiat. Res. 142, 288–294, 1995.
15. Stoll, U., Barth, B., Scheerer, N., Schneider, E., and Kiefer, J., Int. J. Radiat. Biol. 70, 15–22, 1996.
16. Strahlenschutzverordnung, Verordnung über den Schutz vor Schäden durch ionisierende Strahlen, Bundesgesetzblatt der Bundesrepublik Deutschland (BGBl) I,2001, 1714, 2001.
17. Townsend, L. W., Shinn, J. L., and Wilson, J. W., Radiat. Res. 126, 108–110, 1991.
18. UNSCEAR 1993 REPORT: Sources and effects of ionizing radiation, United Nations, New York, 1993.
19. UNSCEAR 2000 REPORT: Sources and effects of ionizing radiation, United Nations, New York, 2000.
20. Wilson, J. W., Health Physics 79, 470–494, 2000.

Index